東北メディカル・メガバンク計画による
健康医療関連ビッグデータの創生

本講演の内容

- ゲノムデータの可能性と限界
- 個別化医療への課題と展望
- 東北メディカル・メガバンク計画によるデータ創生
- ToMMoスーパーコンピュターシステム
- 機微性の高いデータ共有に向けて
- まとめ

2022年12月26日@UDACキックオフシンポ
東北大学
情報科学研究科
東北メディカル・メガバンク機構
木下賢吾

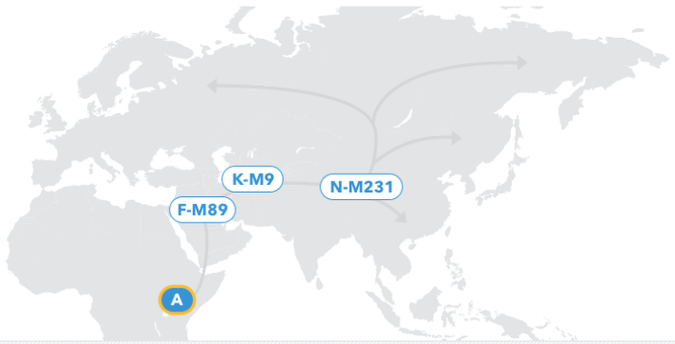
自己紹介

Migrations of Your Paternal Line

A 275,000 Years Ago

Haplogroup A

The stories of all of our paternal lines can be traced back over 275,000 years to just one man: the common ancestor of haplogroup A. Current evidence suggests he was one of thousands of men who lived in eastern Africa at the time. However, while his male-line descendants passed down their Y chromosomes generation after generation, the lineages from the other men died out. Over time his lineage alone gave rise to all other haplogroups that exist today.



F-M89 76,000 Years Ago

K-M9 53,000 Years Ago

N-M231 45,000 Years Ago

N-M231 45,000 Years Ago

Origin and Migrations of Haplogroup N-M231

Your paternal line stems from haplogroup N-M231, the root of all branches of N. The lineage likely originated in southeastern Asia, probably in or around modern-day southern China, between about 38,000 and 45,000 years ago years ago. Much later, perhaps as the Ice Age wound down about 12,000 years ago and glaciers retreated from northern Eurasia, men carried this haplogroup north toward the Altai Mountains and the Tibetan Plateau. After a pause, the haplogroup began expanding about 3,000 years ago into present-day Russia, where it became very common among speakers of the Uralic languages that were developing in the Ural Mountains and Volga River drainage. After another millennium men carried haplogroup N even farther north and west into eastern Europe and Scandinavia. Today N is distributed throughout northern Eurasia, with an extension by some older branches of the haplogroup into southeastern Asia. It ranges from Japanese populations in the far East to Norwegian populations in the far West. It is particularly common among indigenous Siberian populations that speak Uralic languages, such as the Enets and Nenets, and Altaic languages, such as the Evenki and Oroqen.

N-M231 Today

N-M231 is extremely rare among 23andMe customers.

Today, you share your haplogroup with all the men who are paternal-line descendants of the common ancestor of N-M231, including other 23andMe customers.

Fewer than 1 in 300,000

23andMe customers share your haplogroup assignment.

あなたの父系は, Nのすべての枝の根であるハプログループN-M 231から生じる.この系統は**約38,000年前から45,000年前の間に,東南アジア,おそらく現在の中国南部またはその周辺で発生**したと考えられる。はるか後、おそらく約12,000年前に氷河期が終わり、氷河がユーラシア北部から後退したとき、このハプログループの人々がアルタイ山脈やチベット高原に向かって北上したのだろう。しばらくして、ハプログループは**約3,000年前から現在のロシアに拡大**し始め、ウラル山脈やボルガ川流域で発達していたウラル諸語の話者に非常によく見られるようになった。さらに千年後には、ハプログループNはさらに北と西に東ヨーロッパとスカンジナビアに運ばれた。**今日, Nはユーラシア北部全域に分布し,ハプログループの古い枝によって東南アジアに拡大している。極東の日本人集団から極東のノルウェー人集団にまで及ぶ**。特に、エネット語やネネット語などのウラル諸語や、エベンキ語やオロケン語などのアルタイ諸語を話すシベリアの原住民に多くみられます。

Migrations of Your Maternal Line

L 180,000 Years Ago

Haplogroup L

If every person living today could trace his or her maternal line back over thousands of generations, all of our lines would meet at a single woman who lived in eastern Africa between 150,000 and 200,000 years ago. Though she was one of perhaps thousands of women alive at the time, only the diverse branches of her haplogroup have survived to today. The story of your maternal line begins with her.

L3 65,000 Years Ago

M 50,000 Years Ago

D 40,000 Years Ago

D 40,000 Years Ago

Origin and Migrations of Haplogroup D

The common ancestor of haplogroup D was a woman who lived in Asia nearly 40,000 years ago. There are two major branches of the D haplogroup in Asia. D5, which is comparable in age to D itself, is common in southern China but rare farther north. D4, a younger haplogroup that arose about 25,000 years ago, is more common in northern Asia, reaching 18% in southern Siberia.

Haplogroup D4 is particularly common among Koreans and in the populations of Manchuria, which is just north of the Korean Peninsula. Recent archaeological discoveries suggest that the earliest inhabitants of Korea probably came from the Altai-Sayan and Baikal regions of Southeast Siberia. They likely began to move into the region by about 30,000 years ago, when they followed mammoths and other large animals into the peninsula. Among Siberian populations, haplogroup D is most common in the Yupik and Chukchi, two modern indigenous groups in northeastern Siberia whose ancestors are thought to have played a significant role in the peopling of the Americas.

D4a2a < 8,000 Years Ago

Your maternal haplogroup, D4a2a, traces back to a woman who lived less than 8,000 years ago.

That's nearly 320 generations ago! What happened between then and now? As researchers and citizen scientists discover more about your haplogroup, new details may be added to the story of your maternal line.

D4a2a Today

D4a2a is rare among 23andMe customers.

Today, you share your haplogroup with all the maternal-line descendants of the

1 in 25,000

23andMe customers share

ハプログループDの共通の祖先は、**約4万年前にアジアに住んでいた女性**である。アジアのDハプログループには2つの主要な分岐がある。D 5はD自体と同程度の年齢であり,中国南部では一般的であるが,それ以上北部ではまれである。約25,000年前に発生したより若いハプログループである**D 4はアジア北部でよくみられ,シベリア南部では18%に達する。ハプログループD 4は韓国人と朝鮮半島のすぐ北にある満州の住民に特によくみられる**。最近の考古学的発見は,韓国の最も初期の居住者がおそらく南東シベリアのアルタイ・サヤンとバイカル地域から来たことを示唆している。彼らは、マンモスや他の大型動物を追って半島に入った約3万年前までに、この地域に進出し始めた可能性が高い。シベリア人集団の中では,ハプログループDは,祖先がアメリカ大陸の定住に重要な役割を果たしていたと考えられているシベリア北東部の2つの近代的先住民族であるユピック族とチュクチ族で最も一般的である。

もう少し最近の私

- 大阪生まれ@1970年
- 京都大学育ち@1999年ようやく学生を終える
  - ボス：郷信広
  - 物理と化学が基本的なバックグラウンド
- ポスドク@阪大　（2年）
  - ボス：中村春木
- 助教@横浜市大 (3.5年)
  - ボス：木寺詔紀
- 准教授@東大医科研　（5年）
  - ボス：中井謙太
- 教授@東北大情報科学　（since 2009.10）
- 副機構長@メガバンク（since 2016.4）

本日の話題

- ゲノムデータの可能性と限界
- 個別化医療への課題と展望
- 東北メディカル・メガバンク計画によるデータ創生
- 機微性の高いデータ共有に向けて
- ToMMoスーパーコンピュターシステム
- まとめ

本日の話題

- **ゲノムデータの可能性と限界**
- **個別化医療への課題と展望**
- 東北メディカル・メガバンク計画によるデータ創生
- 機微性の高いデータ共有に向けて
- ToMMoスーパーコンピュターシステム
- まとめ

2022年の現実

(12) **United States Patent**
Wojcicki et al.

(10) **Patent No.: US 8,543,339 B2**
(45) **Date of Patent: Sep. 24, 2013**

(54) **GAMETE DONOR SELECTION BASED ON GENETIC CALCULATIONS**

(75) Inventors: **Anne Wojcicki**, Palo Alto, CA (US); **Linda Avey**, Lafayette, CA (US); **Joanna Louise Mountain**, Menlo Park, CA (US); **John Michael Macpherson**, Palo Alto, CA (US); **Joyce Yeh-hong Tung**, Menlo Park, CA (US)

(73) Assignee: **23andMe, Inc.**, Mountain View, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 347 days.

(21) Appl. No.: **12/592,950**

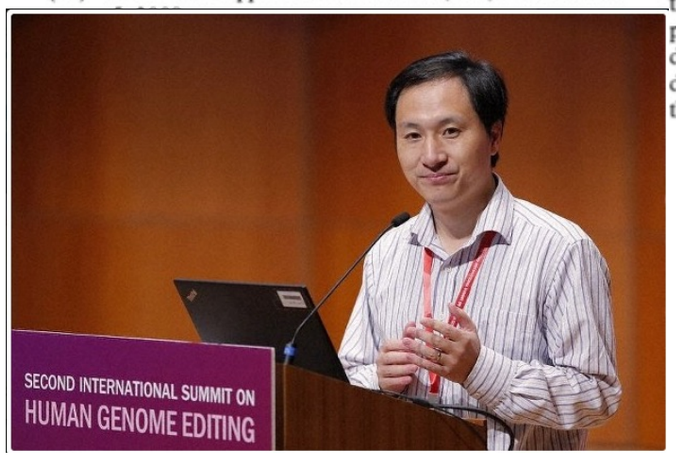
(22) Filed: **Dec. 4, 2009**

(65) **Prior Publication Data**

US 2010/0145981 A1 Jun. 10, 2010

Related U.S. Application Data

(60) Provisional application No. 61/201,101, filed on Dec.



(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,844,156	B2 *	1/2005	Rosen	435/6.18
7,599,802	B2 *	10/2009	Harwood et al.	702/20
7,668,658	B2 *	2/2010	Koster et al.	702/19
2005/0278125	A1 *	12/2005	Harwood et al.	702/20
2006/0008859	A1 *	1/2006	Seul et al.	435/7.25
2006/0205001	A1 *	9/2006	Zhang et al.	435/6
2007/0042369	A1	2/2007	Reese et al.	
2007/0093968	A1 *	4/2007	Zhang et al.	702/19
2009/0299645	A1 *	12/2009	Colby et al.	702/19
2010/0022406	A1 *	1/2010	Srinivasan et al.	506/9
2010/0191735	A1 *	7/2010	Reiss et al.	707/740
2011/0124515	A1 *	5/2011	Silver	506/8
2012/0078901	A1 *	3/2012	Conde	707/736

* cited by examiner

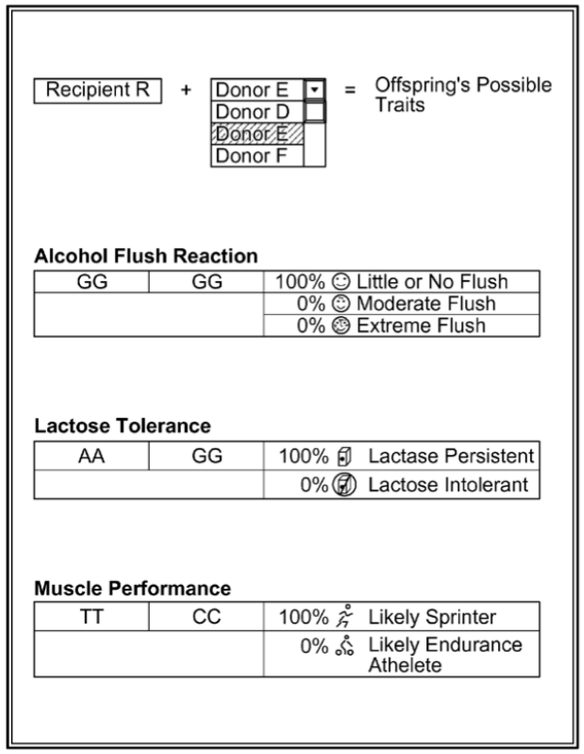
Primary Examiner — Mary Zeman
(74) *Attorney, Agent, or Firm* — Van Pelt, Yi & James LLP

(57) **ABSTRACT**

Gamete donor selection includes receiving a specification including a phenotype of interest, receiving a genotype of a recipient and a plurality of genotypes of a respective plurality of donors, determining statistical information pertaining to the phenotype of interest based at least in part on different pairings of the genotype of the recipient and a genotype of a donor in the plurality of donors, and identifying a preferred donor among the plurality of donors, based at least in part on the statistical information determined.

28 Claims, 7 Drawing Sheets

賀建奎（が・けんけい、He Jiankui）
claimed to have genetically ‘edited twins’
by gene editing to human embryos.
Nov. 26, 2018



2020年1月12日朝日新聞

現時点でのゲノムの持つ可能性

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem ... can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Science, 339, 321-324, 2013

ゲノム情報から公共データを利用して名字を推定

ゲノムから顔貌の予測→犯罪捜査への期待

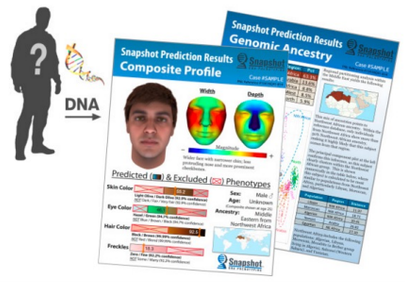
Stranger Visions@Heather Dewey-Hagborg

June 2012

商用サービスも展開

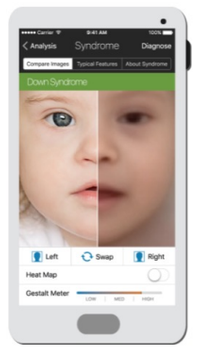
The Snapshot DNA Phenotyping Service

DNA Phenotyping is the prediction of physical appearance from DNA. It can be used to generate leads in cases where there are no suspects or database hits, or to help identify unknown remains.



https://snapshot.parabon-nanolabs.com

逆に顔の形状から遺伝性疾患（ゲノムで決まる希少疾患）の有無を判断するアプリ

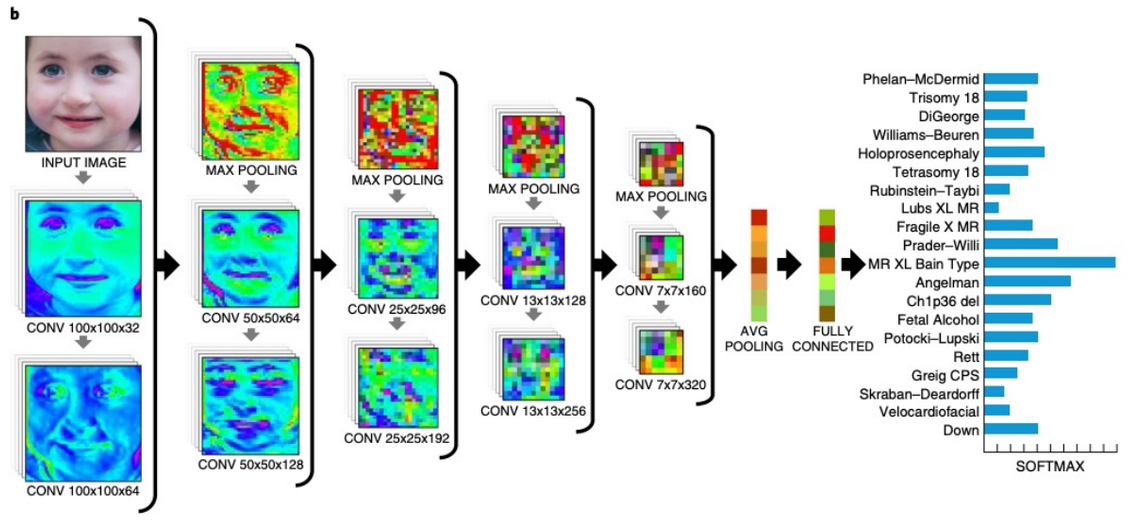


Detect Phenotypes & Reveal Relevant Facial and Non-facial Features

- Detection of phenotypes from facial photos
- Automatic calculation of anthropometric growth charts
- Suggestion of likely phenotypic traits to assist in feature annotation

An objective computer-aided dimension to the art of dysmorphology

Dr. Michael Hayden, Clinical Genetics



Identifying facial phenotypes of genetic disorders using deep learning

Nature Med, 2019

Yaron Gurovich[1]*, Yair Hanani[1], Omri Bar[1], Guy Nadav[1], Nicole Fleischer[1], Dekel Gelbman[1], Lina Basel-Salmon[2,3], Peter M. Krawitz[4], Susanne B. Kamphausen[5], Martin Zenker[5], Lynne M. Bird[6,7] and Karen W. Gripp[8]

現時点でのゲノムの持つ可能性その2

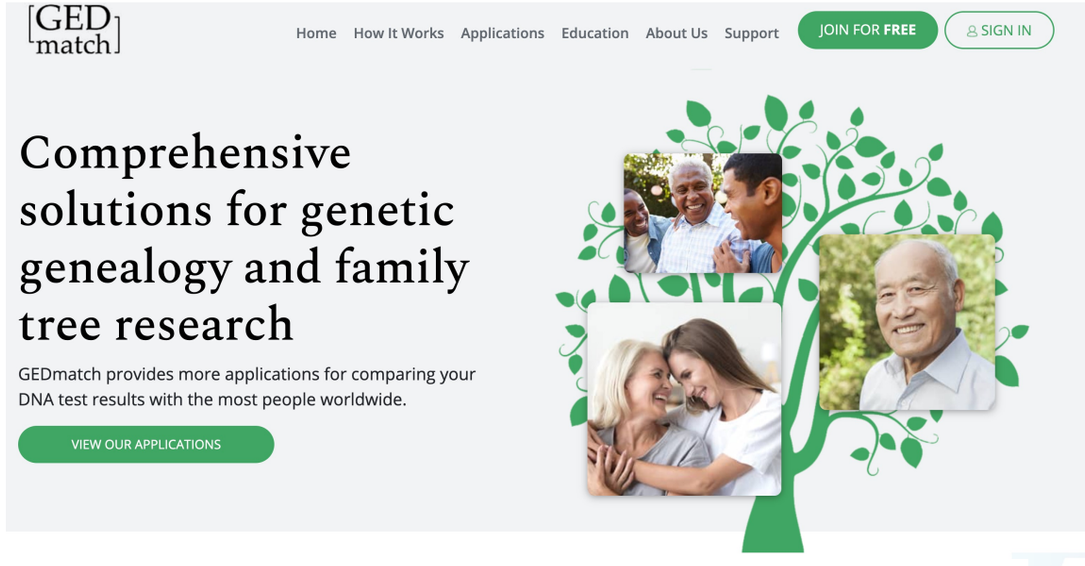
2018年4月28日の現実

DNA analysis site that led to the Golden State Killer issues a privacy warning to users

Taylor Hatmaker @tayhatmaker / Apr 28, 2018 Comment

As more details emerge about the arrest of the man suspected to be the Golden State Killer, it's clear that one of the most infamous unsolved cases of all time was cracked using a popular free online genealogy database.

The site, known as GEDmatch, is a popular resource for people who have obtained their own DNA through readily available consumer testing services and want to fill in missing portions of their family tree to conduct further analyses. Compared to a polished service like 23andMe, GEDmatch is an open platform lacking the same privacy and legal restrictions that govern user data on more mainstream platforms.



How it Works?

1. Take a DNA test and download the results as a DNA data file.
2. Upload the DNA data file to GEDmatch for processing.
3. Explore matching and comparison reports and other DNA tools.

name E-mail

Select	Match No.	Kit	Name (* => alias)	Email	GED WikiTree	Age(days)	Type	Sex	Haplogroup Mt	Haplogroup Y	Autosomal Total cM	Autosomal Largest	Autosomal Gen	X-DNA Total cM	X-DNA Largest	Source	Overlap
	1	FU7915582	[redacted]	[redacted]		1276	2	M			35.8	11.7	4.32	0	0		76974
	2	PP6333198				380	2	M			27.2	11.5	4.52	0	0	23andMe	73948
	3	WW6420450				1645	2	F			23	15.6	4.64	0	0	23andMe	83702
	4	YY9606559				1737	2	F	A5c		22.1	14.8	4.67	0	0	23andMe	82522
	5	FF2500220				1890	2	F			21.5	12.6	4.69	0	0	23 and me	83544
	6	DT5967798				1710	2	F			20.9	11.1	4.71	0	0	23andMe	83537
	7	WX4824288				962	2	M			20.9	13.6	4.71	0	0	GESE	76516
	8	JR8770841				599	2	M	C	T-L208	20.8	20.8	4.71	0	0	23andMe	78081
	9	QE1931184				1641	2	F	R9b1b		20.4	11.3	4.73	0	0	23andMe	83759
	10	UF6217757				849	2	F	Y1		20.3	11.1	4.73	0	0	23andMe	83643
	11	GV6369005				625	2	M			20.2	12.7	4.74	0	0	23andMe	83990
	12	ZS9617780				9	2	F			19.8	10.2	4.75	0	0	23andMe	83642
	13	BL2380168				1363	2	F	Y1		19.7	10.8	4.75	0	0	23andMe	83642
	14	SE7227864				1060	2	F	A3		19.7	11.1	4.76	0	0	23andMe	84144

身近になったゲノム情報

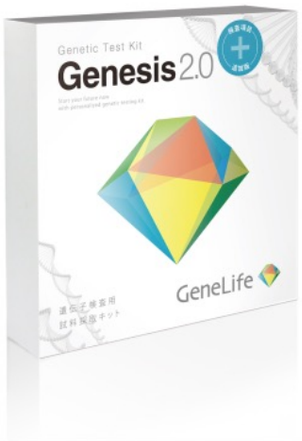
ジーンライフジェネシス2.0プラス

Genesis2.0

検査項目 追加版

肥満タイプや疾患のリスクなど約360項目を解析できる遺伝子検査キット。大腸がんや乳がん、心筋梗塞などの疾患リスクと、肥満タイプなどの体質を手軽に分析できる検査です。

対象年齢 20才～　検査期間 約1ヶ月　発送目目安 5営業日以内



Genesis2.0に検査項目を追加し、お求めやすい価格で新登場

通常価格 ¥14,900（税込）

カートに入れる

新しく追加された検査項目

COVID-19重症化

コロナ禍において新しい検査項目追加と予防に関する注意事項の追加

COVID-19重症化リスクを評価する新しい遺伝子検査項目の追加

感染症対策を行いつつ、あなたの遺伝子からも学びましょう

- 遺伝的体質と生活習慣病
- 食事とBMI
- 運動とスポーツ

23andMe

Find out what your DNA says about you and your family.

- See how your DNA breaks out across 2000+ regions worldwide
- Discover DNA relatives from around the world
- Share reports with family and friends
- Learn how your DNA influences your facial features, taste, smell and other traits

order now　USD$99

病気と体質に関する全検査（280項目）が入ったセット

病気（3大疾病のがん・心筋梗塞・脳梗塞 等）と体質（長生き・肥満・肌質 等）280項目に加え、新型コロナウイルス関連項目の遺伝的傾向がわかるスタンダードパッケージ

※当検査項目はがんパックの内容を全て含んでいます

価格：32,780円(税込)

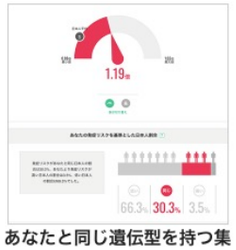
この検査メニューを申し込む

遺伝子検査結果（レポート）は、検査完了後、MYCODEサイト上でご覧いただけるようになります。
充実の内容を、サンプルでご紹介します。

病気・体質の検査結果レポート

遺伝子の知識がなくても理解しやすい解説付きレポート

- あなたの遺伝型を解析し、病気の発症リスクや体質の遺伝的な傾向についてお伝えします。
- 各項目ごとに、病気や遺伝子についての知識がなくても理解しやすい詳しい解説が付いています。



あなたと同じ遺伝型を持つ集団のリスクを表示

結果だけでなく、検査項目の意味も詳しく説明

解析した遺伝型や参考論文を公開

気になる項目は「MYボックス」で管理

COVID-19へのリスク

あなたの遺伝型を見る >
採用論文を見る >

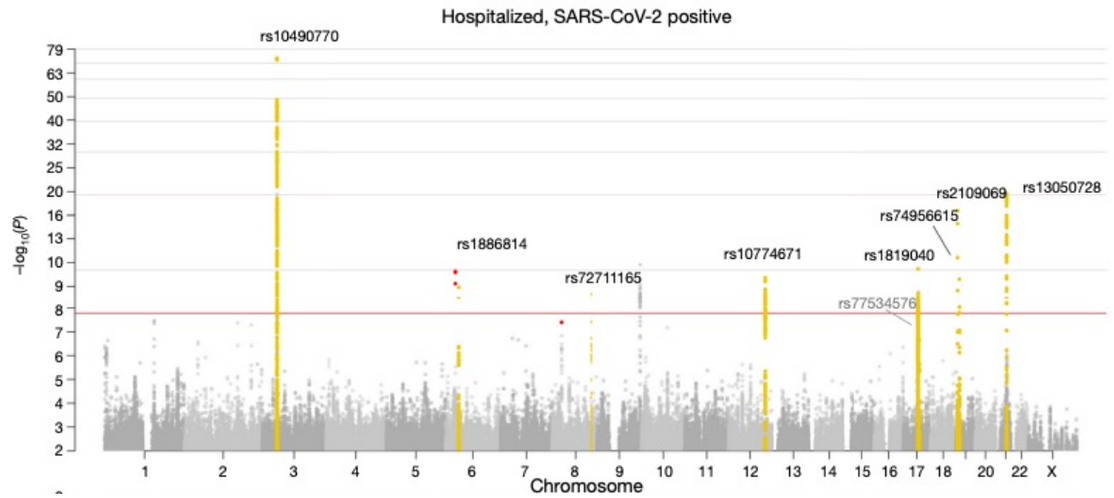
あなたの遺伝型における、新型コロナウイルス感染時の重症化リスク(OAS1遺伝子に基づく)は日本人平均の1.07倍でした。 ?

あなたの重症化リスクを基準とした日本人割合

低い	同じ
30.1%	70.1%

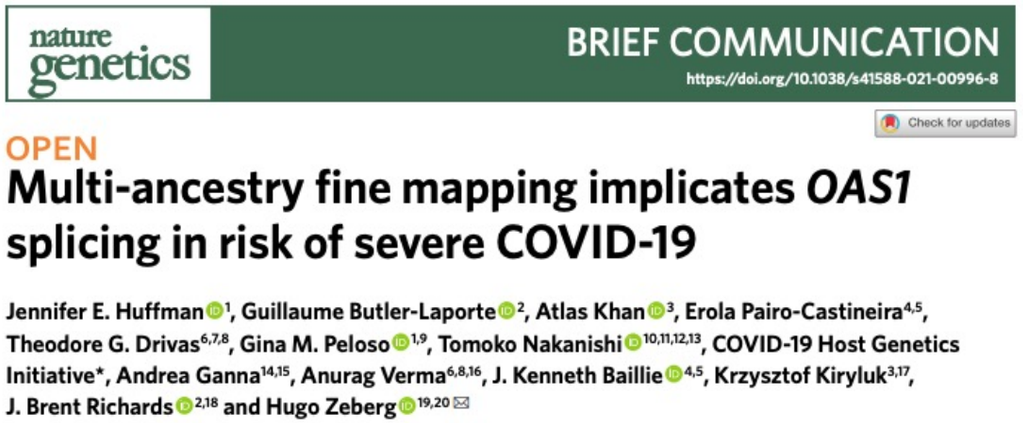
※日本人割合は、Hapmapの「日本人の頻度データ」を利用して算出しています。小数点以下を四捨五入しているため、合計が100%にならない可能性もあります。

Nature, Dec, 2021, https://doi.org/10.1038/s41586-021-03767-x



新型コロナウイルス感染症の重症化リスクの項目追加にあたり、MYCODEではThe COVID-19 Host Genetics Initiativeにより2021年7月に科学雑誌「Nature」で発行された研究論文である、「Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis」を採用論文としました。
論文で新型コロナウイルス感染症の重症化リスクに関連すると報告されたSNPの中から、MYCODEの基準を満たすと判断されたOAS1遺伝子領域に存在するrs10774671について、COVID-19疾患の重症化リスクにかかわる遺伝型としてお調べしております。
【OAS１遺伝子について】
OAS1遺伝子は12番染色体上に存在し、2',5'-オリゴアデニル酸合成酵素というタンパク質の設計図となる遺伝子で、ウイルス感染に対する自然免疫応答に関連があると考えられています。

なお、**2022年2月にはOAS1の近くのrs10774671が原因とは考えにくいという論文も出ている**


nature genetics
BRIEF COMMUNICATION
https://doi.org/10.1038/s41588-021-00996-8
Check for updates
OPEN
Multi-ancestry fine mapping implicates *OAS1* splicing in risk of severe COVID-19
Jennifer E. Huffman[1], Guillaume Butler-Laporte[2], Atlas Khan[3], Erola Pairo-Castineira[4,5], Theodore G. Drivas[6,7,8], Gina M. Peloso[1,9], Tomoko Nakanishi[10,11,12,13], COVID-19 Host Genetics Initiative*, Andrea Ganna[14,15], Anurag Verma[6,8,16], J. Kenneth Baillie[4,5], Krzysztof Kiryluk[3,17], J. Brent Richards[2,18] and Hugo Zeberg[19,20]

疾患リスク予測時代の到来

TMMが当初から進めているアレイを使って遺伝子から一人一人の病気になりやすさを予測し、医療の枠組みで活用する構想は、最近、世界の潮流となりつつある

NATURE MEDICINE | VOL 24 | OCTOBER 2018 | 1483 |

editorial

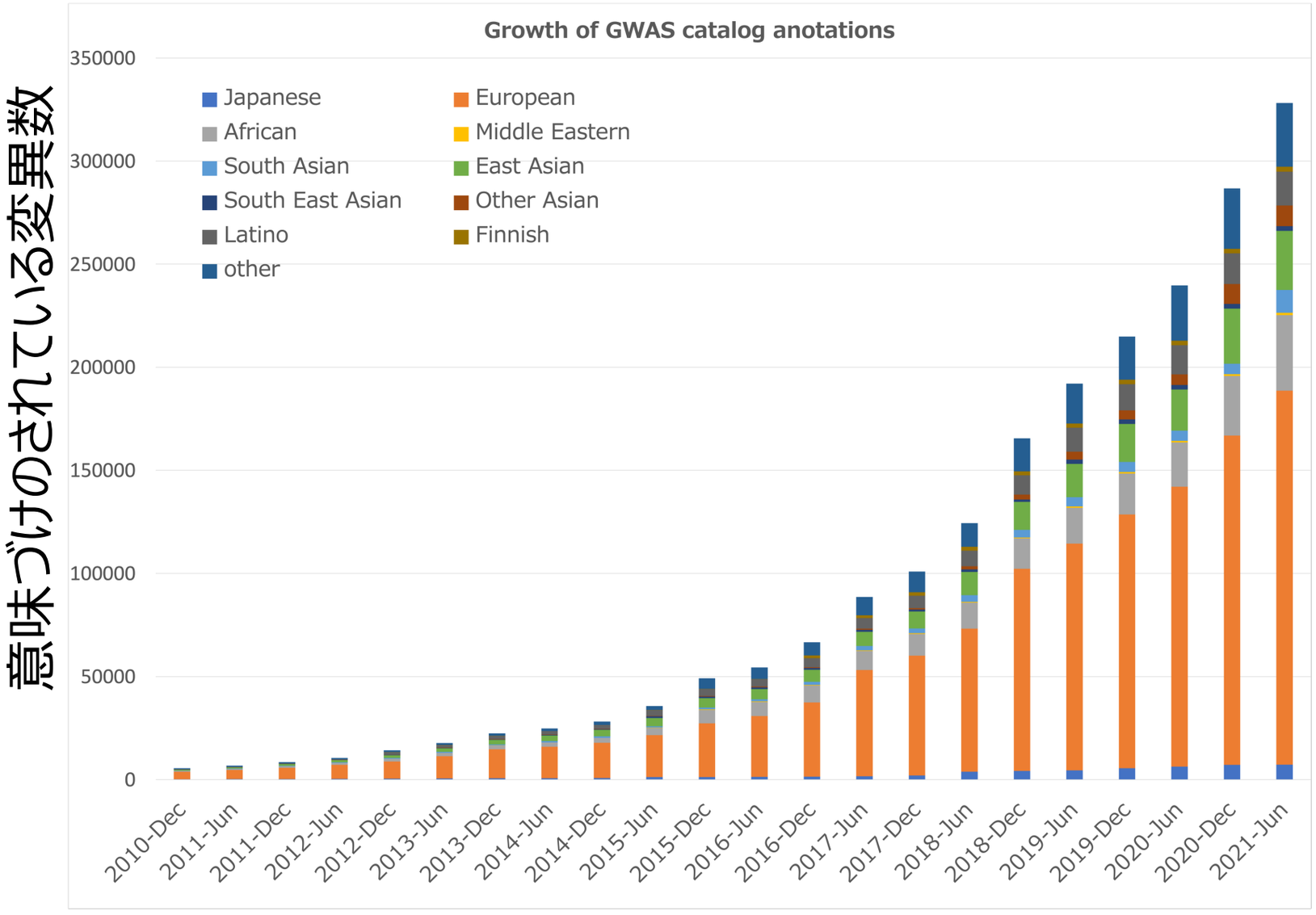
GWAS to the people

Thanks to improvements in data collection and analysis, some polygenic risk scores that predict disease risk are approaching the same predictive accuracy offered by tests for monogenic mutations. The time to think about how best to incorporate polygenic tests in the clinic is now.

Polygenic risk score (PRS；多遺伝子リスクスコア)

- 遺伝子バリアントを使って個人が複雑な遺伝性疾患を発症する確率を計算する**多遺伝子リスクスコア（PRS）**の研究は大きく進展し、現在では**多因子疾患について発症リスクを予測できる信頼性の高いスコアが比較的容易に得られるようになりつつある**
- 一般的な集団よりも心血管疾患を発症する可能性が高いことを示す多遺伝子リスクスコア（PRS）が示されたら、それは医者の指導に従って生活スタイルを変えたり、コレステロールを下げる薬を処方してもらったりする動機になるだろう
- **全ゲノム塩基配列解読**はやはりコストが高いが、**それに比べてPRS算出に使われるアレイチップは一般に100ドル以下であり使いやすい**
- 克服すべき課題はまだまだ多くあるが、**ゲノム情報を基盤として健康状態を理解する方策は、全ての人が使えるようにするべきであり、そのための検査を臨床現場に導入する最良の方法を考えなければならない時期が来ているのである**

日本人におけるエビデンスの不足



東アジア系10%

メインは欧州系

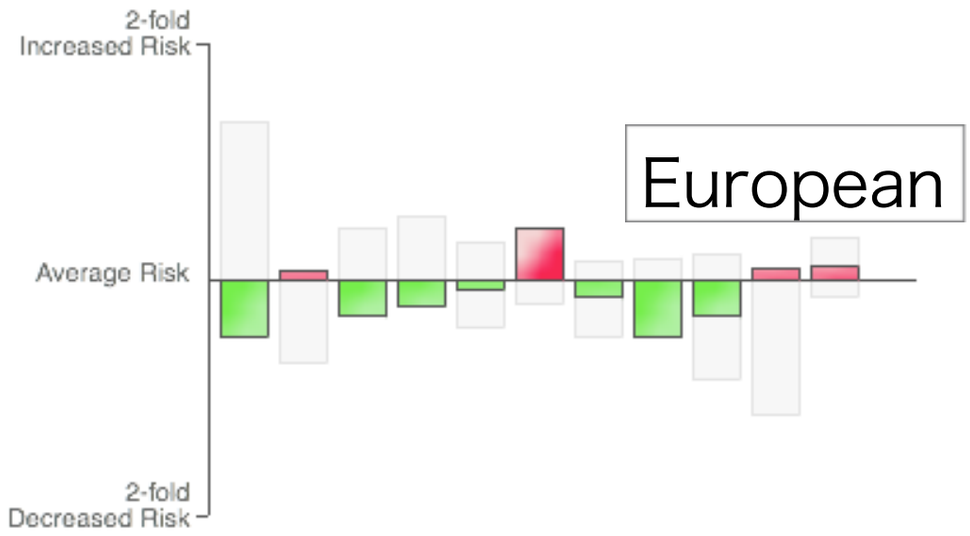
日本人は3.9%

- ほとんどの先行研究が欧米主体である
- 同じ疾患でも日本人と欧米人では関連変異が違うことも多い
- ゲノム解析は簡便になってきたが、日本人に裨益していない

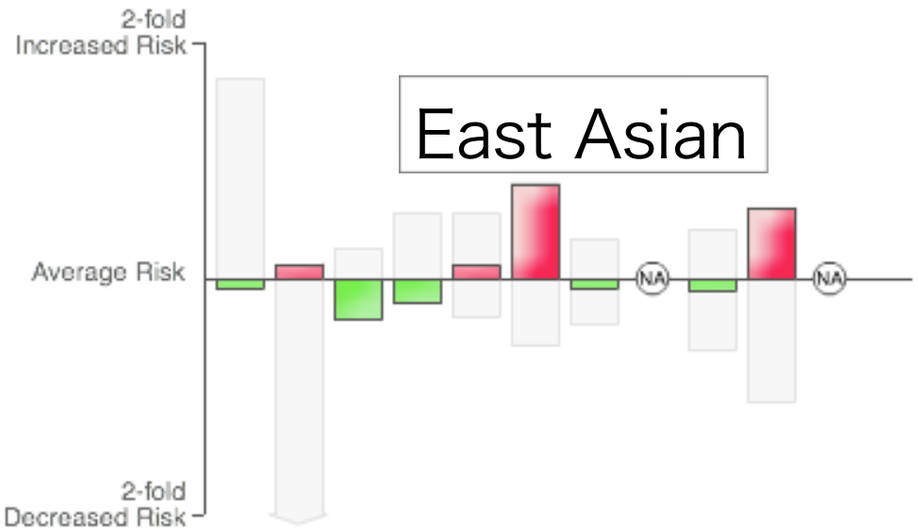
日本人特有の変異・疾患ＤＢの必要性

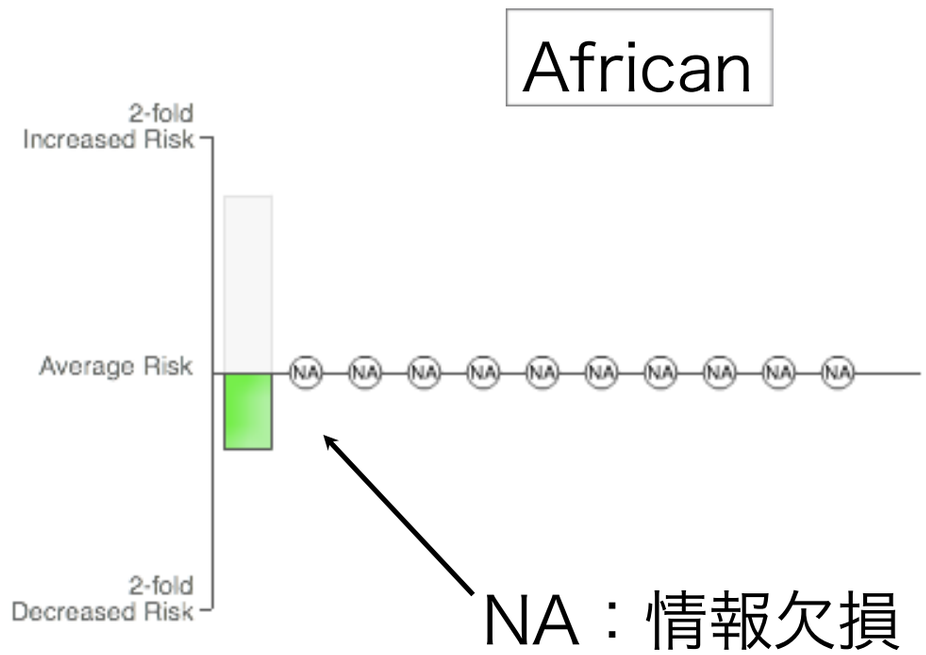
同じ変異でも民族によって遺伝的リスクは異なる
情報が無い変異の解析は非常に困難

民族を考慮した，Type 2 Diabetesのrisk算出例(23andMeより)



・あるヒトのgenotypeからriskを算出
・各barは同じマーカーSNP

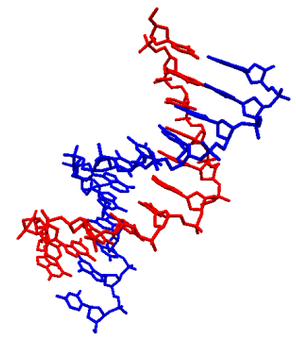




遺伝・環境相互作用：ゲノムだけでは決まらない

遺伝要因と環境要因

遺伝要因



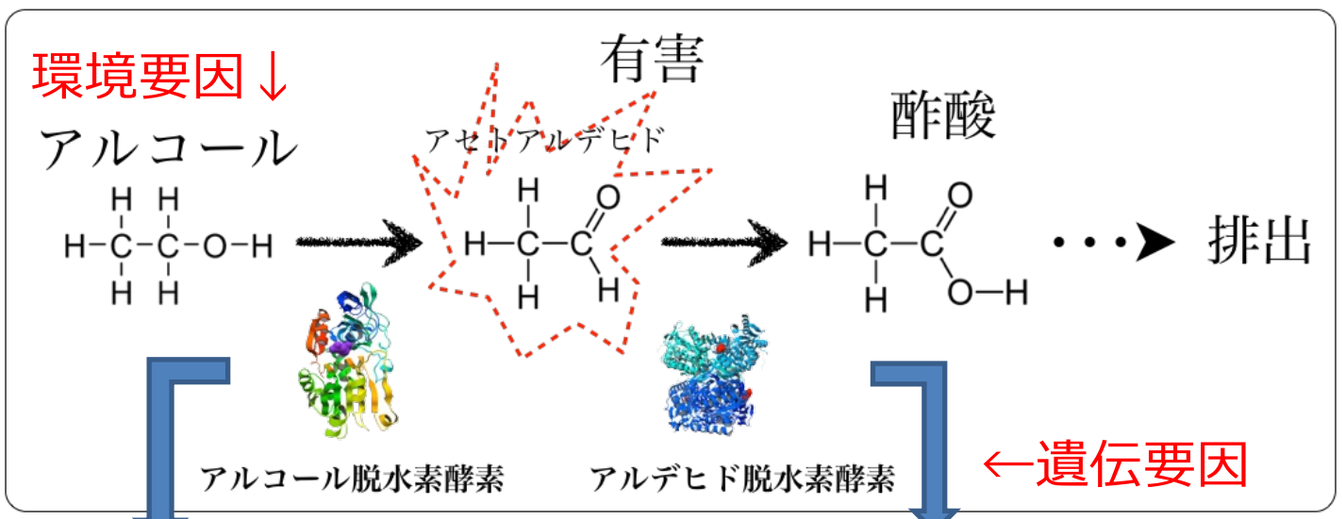
遺伝子だけで
ほぼ決まる形質
単一遺伝子疾患

遺伝子x環境で決まる形質
身長、体重
多因子疾患、生活習慣病

環境だけでほぼ決まる形質
雪で転んですりむいた傷

環境要因
食事、運動
飲酒、喫煙
天候など

1 環境要因 x 2 遺伝子 = お酒との関係



AUGAGCACAGCAGGAAAAGUAAUCAAAUGCAAAGCAGCU
GUGCUAUGGUAGUAAAGAACCCUUUUCCAUUGAGGAU
GUGGAGGUGCUGUAAGGCUUAUGAAGUUCGCAUU
AAGAUGGUGCUGUAGCUCUGUCGCACAGAUGACCAC
GUGGUUAGUGGCAACCUGGUGACCCCCCUUCCUGUGAUU
UUAGGCCAUGAGCAGCCGGCAUCGUGGAGAGUGUUGGA
GAAGGGGUUAUAGGUAACCAGGUGAUAAAGUCAUC
CCGCUCUUUACUCCUCAGUGUGGAAAAUGCAGAGUUUGU
AAAAACCCGGAGAGCAACUACUGCUUGAAAAAUGAUCUA

A型：CAC
G型：CGC
GG 64.5%
AG 31.2%
AA 4.0%

GGCUACAUCAACACGGGGAAGCAAGAGGGGCGAAGCU
AUUGCUGCUGACCGUGGUUACUGCCAGCCACUGU
GAUGGCAUGACCAUCGCCAAGGAGGAUGGCGC
AAGUUCAAGACCAUAGAGGAGGUUGUGGAGAGCCA
CUGGCCGCAGCUGUCUUCACAAGGAUUGGACAAGG
GCCCUCCAGGCGGGCACUGUGUGGGGAUGA
UCACCCUUUGGUGGCUACAAGAUCGGGGGCCC
GGGCUGCAGGCAUACACUGAAGUGAAAACUGUCACAG
AACUCAUAAGAAUCAUGCAAGCUUCCUCCCUCAGCCA

A型：AAG
G型：GAG
AA 58.2%
AG 35.9%
GG 5.8%

本日の話題

- ゲノムデータの可能性と限界
- 個別化医療への課題と展望
- **東北メディカル・メガバンク計画によるデータ創生と活用**
- **機微性の高いデータ共有に向けて**
- **ToMMoスーパーコンピュターシステム**
- まとめ

日本人のためのゲノム医療実現に必要な研究の方向性

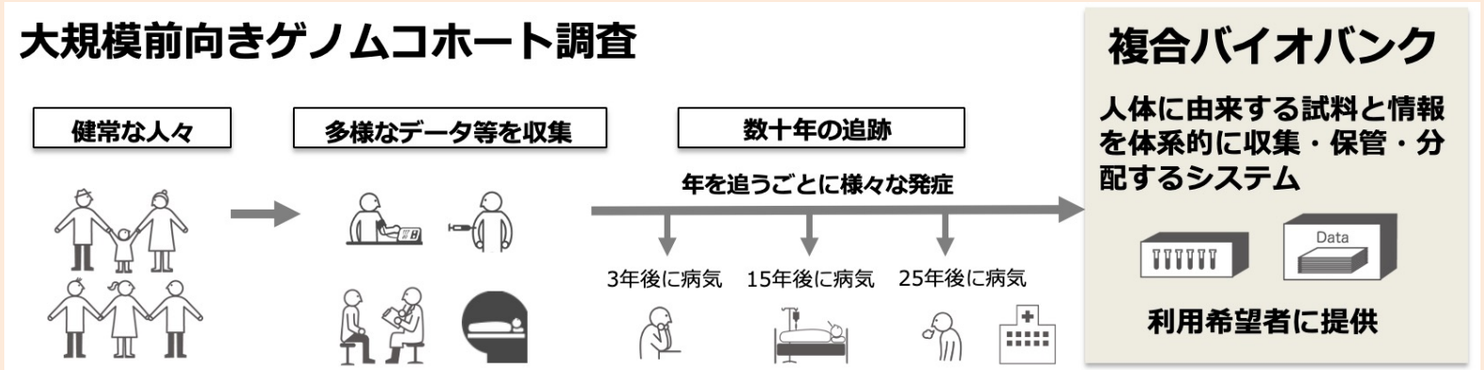
Missing Heritability（失われた遺伝率）の克服が重要

遺伝子変異と疾患を繋ぐエビデンスが圧倒的に不足している

Missing Heritabilityを克服するためには
正確な遺伝要因と環境要因の収集・解析が必要

具体的には、前向きゲノムコホート研究に以下の要素を盛り込むことが重要

- **家系情報**
- **サンプル数の増加**
- **全ゲノム解析とオミックス解析**
- **人生初期からの環境要因把握**
- **正確な表現型の把握**
- **変異と環境の統合解析**
- **他遺伝子の効果を考えたモデル化**
- **医療ビッグデータの利活用**



DNA,血漿、血清、尿など
既往歴など多様な健康調査

発症前後の比較が可能
病気にならなかった人の
データも分かる

生体試料・解析情報を
バンキング

世界の主なコホート・バイオバンク

コホートの規模と家系情報の価値

世界的に大規模家系情報付コホート・バンクに期待が集まっている

- 大規模化により、多くの要因が関わる疾患（多因子疾患）（例：認知症、心血管障害など）や同じ病名の中での多様性（例：糖尿病）に対応できる
- 家系情報により、遺伝子が疾患にどれくらい関係しているかを精確に算出できる

①患者コホート・バンク	②前向きコホート・バンク	③家系情報付前向きコホート・バンク
KIバイオバンク (スウェーデン) 患者 28万人　2003年～	UKバイオバンク（英） 住民 50万人　2006年～	deCODE (アイスランド) 家系付コホート27万人 1998年～（最も成功した例といわれる）
バイオバンク・ジャパン(BBJ　日) 患者 20万人　2003年～	Taiwan Biobank (台) 住民 20万人　2005年～	LifeLines (オランダ) 非妊婦三世代 16万人　2007年～
6NC Biobank (日) 患者・リクルート中　2013年～	Precision Medicine Initiative (米：NIH) 住民 100万人　2015年～	TMM（日本） 東北メディカル・メガバンク計画 ② 住民8万人 ＋ ③ 出生三世代7万人 2012年～ （2種類のコホートを戦略的に実施）

- 患者コホート・バンクは古くより多数存在
- 歴史的には①⇒②⇒③と開発されてきた
- 前向きコホート・バンクは超大規模化

妊婦さんから始めることで、胎児期を含めた生涯の環境データが蓄積できる

全ゲノム解析を実施している世界の代表的な前向きコホート

UK Biobank

- 50万人の40-64歳のボランティア参加者
- 詳細なベースライン検査
- MRイメージング
- 20万人の WGS (Feb 2022)
- オミックス解析に挑戦中
- 研究のための強力なデータベース
- Cloud-based のプラットフォーム

All of US

- 100万人の多彩な参加者をリクルート
- eConsent (電子的な同意取得)
- 参加者による調査票記入情報がある
- Electronic health records
- Fitbit data
- 10万人の WGS (March 2022)
- Cloud-based platform “Researcher Workbench”

TMM

- 15万人の一般住民参加者
- ２つのコホート研究：“CommCohort” & “BirThree cohort”
- 詳細なベースライン検査と追跡調査・家系情報
- MRイメージング
- 5万人の WGS (June 2022)
- オミックス解析に挑戦中
- 複合バイオバンク (Integrated biobank) として試料と情報を分譲中
- Cloud-based のプラットフォームで Data Visiting を実践中

地域住民コホート・三世代コホート

TMM計画では２種類のコホートを活用することにより大きな成果を目指す

地域住民コホート

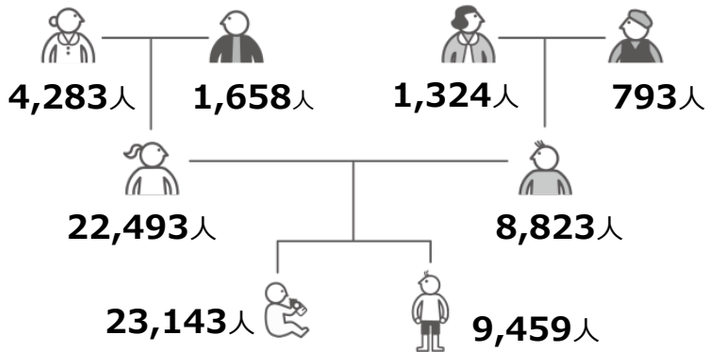
沿岸部を中心に**８万人以上**の成人

宮城登録者　　52,232 名
岩手登録者　　31,861 名
総計　　84,093 名
2016年3月末で新規リクルート完了
目標達成

三世代コホート

産院等で妊婦さんを中心に、子世代、親世代
祖父母世代の三世代、**７万人規模**

登録者　73,529 名*　（2019年8月31日現在）



*曽祖父母78人と拡大家族1,475人を含む

総計15万人以上のリクルート達成

15万人超の参加者に順次、再来所を依頼

TMM計画コホート調査の調査項目

採血： 協力者全員より34㎖の採血

検査項目

採血検査	末梢血一般
	血液像
	血糖
	HbA1c
	GOT
	GPT
	γGTP
	総コレステロール
	HDLコレステロール
	中性脂肪
	尿素窒素
	Cr（eGFRとして回付）
	尿酸
	血清ペプシノゲン
	ヘリコバクターピロリ
	グリコアルブミン
	特異的IgE（5項目）
	総IgE
	シスタチンC

他に、尿・歯垢、唾液、母乳なども採取

調査票による生活習慣等の把握

・標準的な調査項目
（運動、飲酒、喫煙、食事、診療情報、人間関係、女性の健康に関する項目、住所氏名等）
・震災関連項目
（抑うつ、被災状況、ストレス）
・ゲノム関連項目
（体質、出生地等）

参加者の健康づくりに役立つことが明らかになっている項目について、検査結果を回付中

地域支援センターにおける詳細検査

特に、身体年齢を調べる検査を実施（希望者のみ）
眼科的検査（眼底・眼軸長・眼圧・網膜断層写真）、**MRI検査**
聴力検査　呼吸機能検査　家庭血圧　口腔内診察　頸動脈エコー検査
体組成計　踵骨骨密度　脚伸展力検査　タブレットアンケート調査　など

追跡調査*

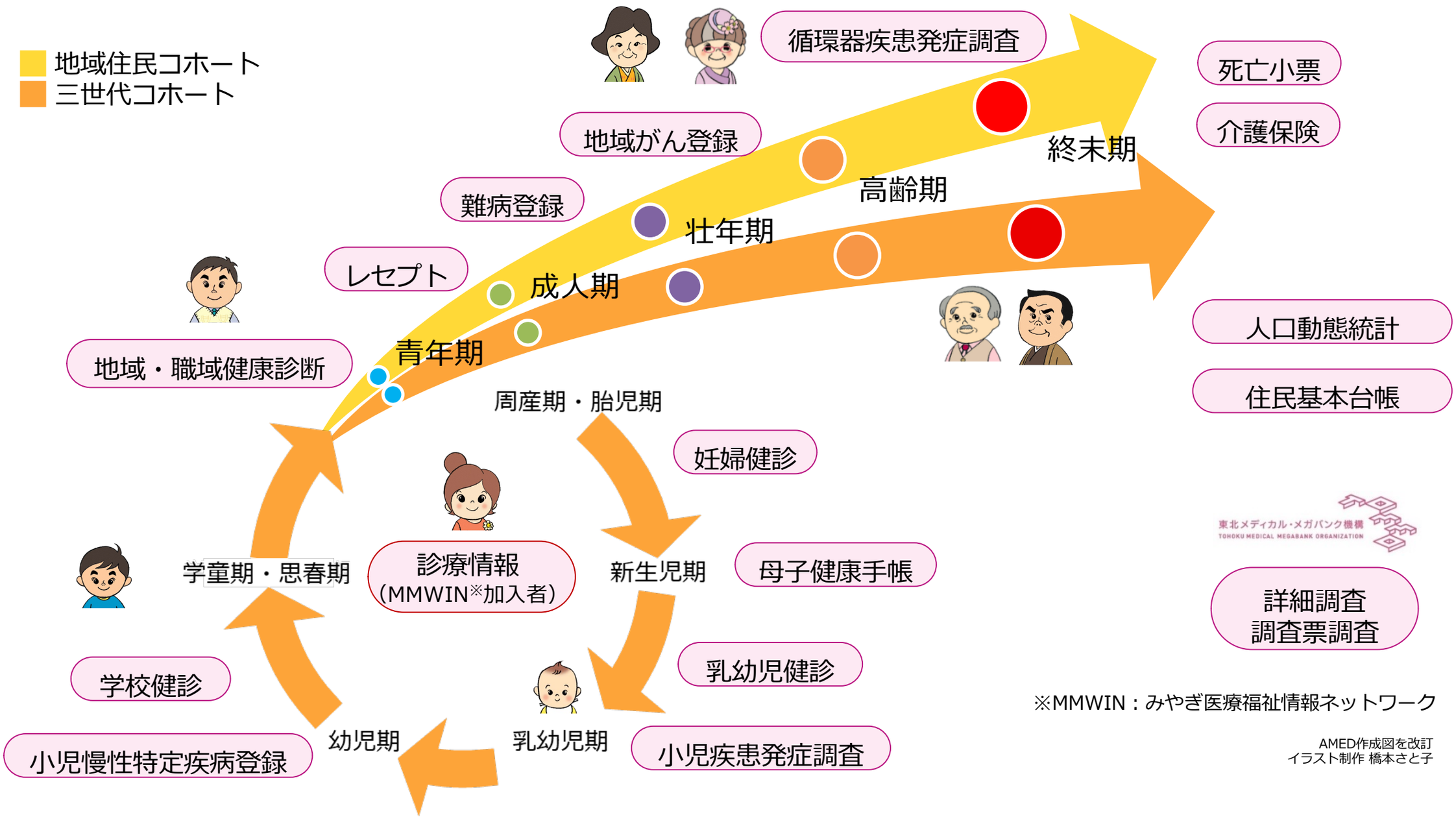
1）調査票による追跡（郵送・Web）/ 2）医療情報活用
3）公的データ・発症登録
4）対面型調査（詳細二次調査）

- 地域支援センターに来所いただき、生理機能検査、バイオバンク用の試料取得（採血）等を行う
- 企業等の協力を得てアドオン（追加）コホート調査を実施する

*以上の追跡調査については参加者からの同意を得ており、2017年度より本格的に実施している

ライフコースデータの情報収集

ライフコースデータのリンケージにより詳細な縦断的解析が可能に



※MMWIN：みやぎ医療福祉情報ネットワーク

AMED作成図を改訂
イラスト制作 橋本さと子

東北メディカル・メガバンク計画の解析戦略

TMMのコホートデザインの特徴

地域住民コホートを基盤とした全ゲノム解読とそれに基づくエスニックアレイ作出、同アレイを用いた全ゲノム解析に三世代コホートを用いた家系解析を組み合わせて、疾患関連遺伝子の同定と検証を目指す先進モデルである

アイスランドdeCODE ジェネティクスの特徴

全ゲノム解読とそれに基づくアレイ解析に広範な家系図を用いた解析を組み合わせて、次々と疾患原因遺伝子を特定している先進モデルであるが、企業が実施している点での限界もある

地域住民コホート

数千人の全ゲノム解析によるリファレンスパネル作製
被災地住民の長期健康調査
環境要因同定

38KJPNまで作成公開

日本人ゲノムリファレンスパネル

疾患NGS解析のフィルターの役割
日本人のアレル構成解析
エスニックアレイの作出
(Japonica Array NEO)

ヘプタファミリーの
全ゲノム解析を完了/分譲中

ジャポニカアレイ

遺伝子型インピュテーションにより全ゲノム補完解析
多くのコホートへの適用して大規模データを得る
100万人コホートなどはNGSをしなくてもOK

15万人解析完了

三世代コホート（再構成された大規模家系）

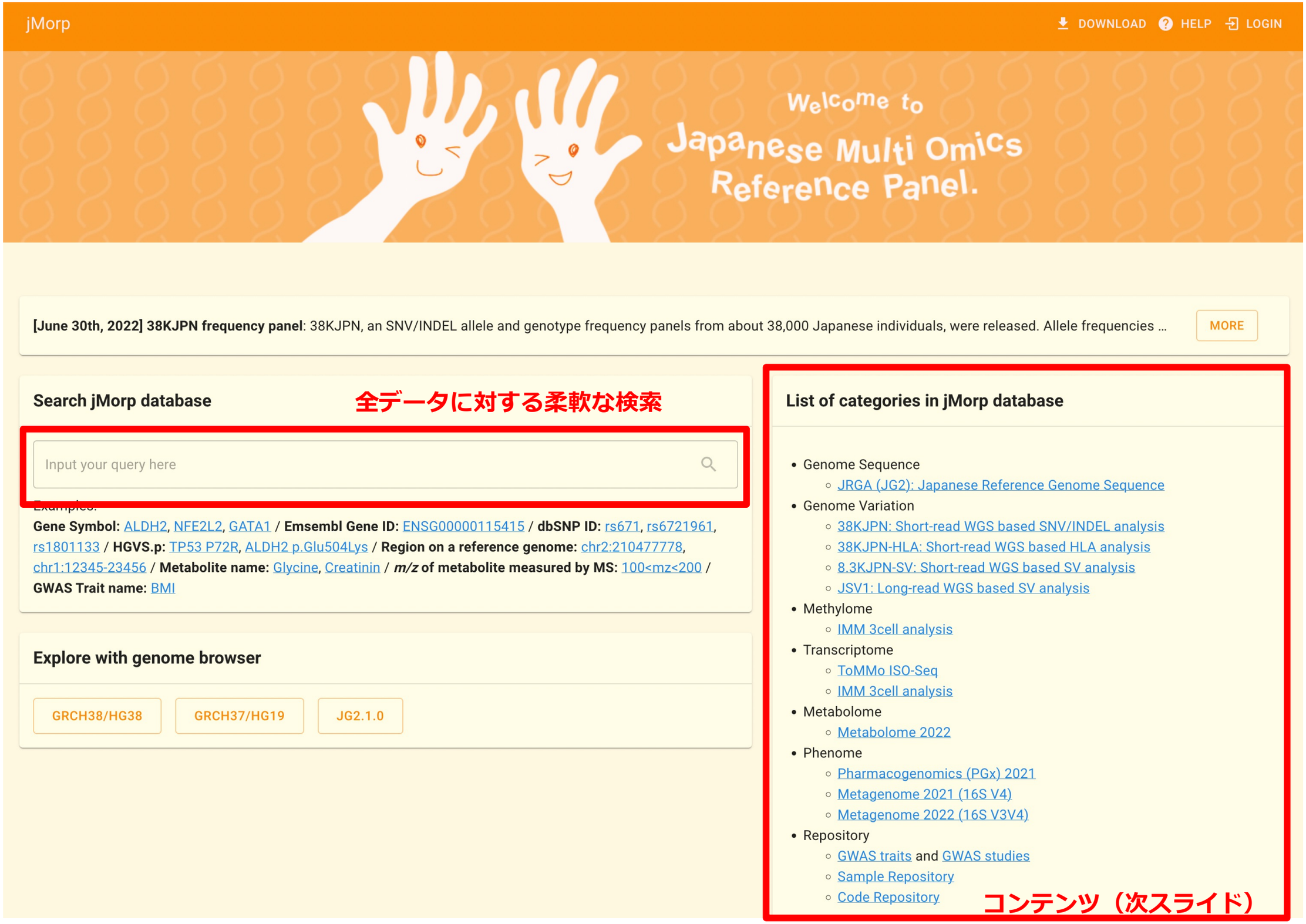
トリオ解析など家系情報を利用した疾患関連遺伝子の絞り込み
De novo変異の解析
産科・小児疾患への取り組み

日本人基準ゲノム
JG2.1公開
JSV1.0公開

アソシエーション解析

遺伝子-環境相互作用の解明

Open Data from jMorp



2022/09/29 renewal

データカテゴリ

List of categories in jMorp database

• Genome Sequence	
◦ JRGA (JG2): Japanese Reference Genome Sequence	日本人男性 3 人の de-novo アセンブリから構築された日本人参照配列
• Genome Variation	
◦ 38KJPN: Short-read WGS based SNV/INDEL analysis	約 38,000 人の日本人短鎖全ゲノム解析から得られた SNV/INDEL アレル頻度・ジェノタイプ頻度データ
◦ 38KJPN-HLA: Short-read WGS based HLA analysis	約 38,000 人の日本人短鎖全ゲノム解析から得られた HLA アレル頻度データ
◦ 8.3KJPN-SV: Short-read WGS based SV analysis	約 8,300 人の日本人短鎖全ゲノム解析から得られた構造多型アレル・ジェノタイプ頻度データ
◦ JSV1: Long-read WGS based SV analysis	222 人の日本人長鎖全ゲノム解析から得られた構造多型アレル・ジェノタイプ頻度データ
• Methylome	
◦ IMM 3cell analysis & coord blood	約 100 人 x 3 種類の血液細胞 の DNA メチル化情報と遺伝子発現情報・アレル頻度情報
• Transcriptome	
◦ ToMMo ISO-Seq	日本人男性 3 人の長鎖トランスクリプトーム解析結果
◦ IMM 3cell analysis & coord blood	約 100 人 x 3 種類の血液細胞 の DNA メチル化情報と遺伝子発現情報・アレル頻度情報
• Metabolome	
◦ Metabolome 2022	約 53,000人の日本人血漿サンプルのメタボローム解析データ
• Phenome	
◦ Pharmacogenomics (PGx) 2021	薬物感受性に関連する酵素における遺伝的多型と酵素活性
◦ Metagenome 16S-v4 2021	歯垢・舌苔サンプルのマイクロバイオーム解析データ(16S v4領域解析)
◦ Metagenome 16S-v3v4 2022	歯垢・舌苔サンプルのマイクロバイオーム解析データ(16S v3v4領域解析)
• Repository	
◦ GWAS traits and GWAS studies	TMMデータを用いた各種GWAS解析
◦ Sample Repository	全ゲノム解析検体の各種解析状況のリポジトリ
◦ Code Repository	コードリポジトリ

ゲノム変異を軸として異なるカテゴリーデータを横断検索可能

38KJPN: 日本人全ゲノム参照パネル
New version of Japanese Genome Reference Panel

50,792検体

2022年1月 data freeze

検体ごとのQC等：-447検体

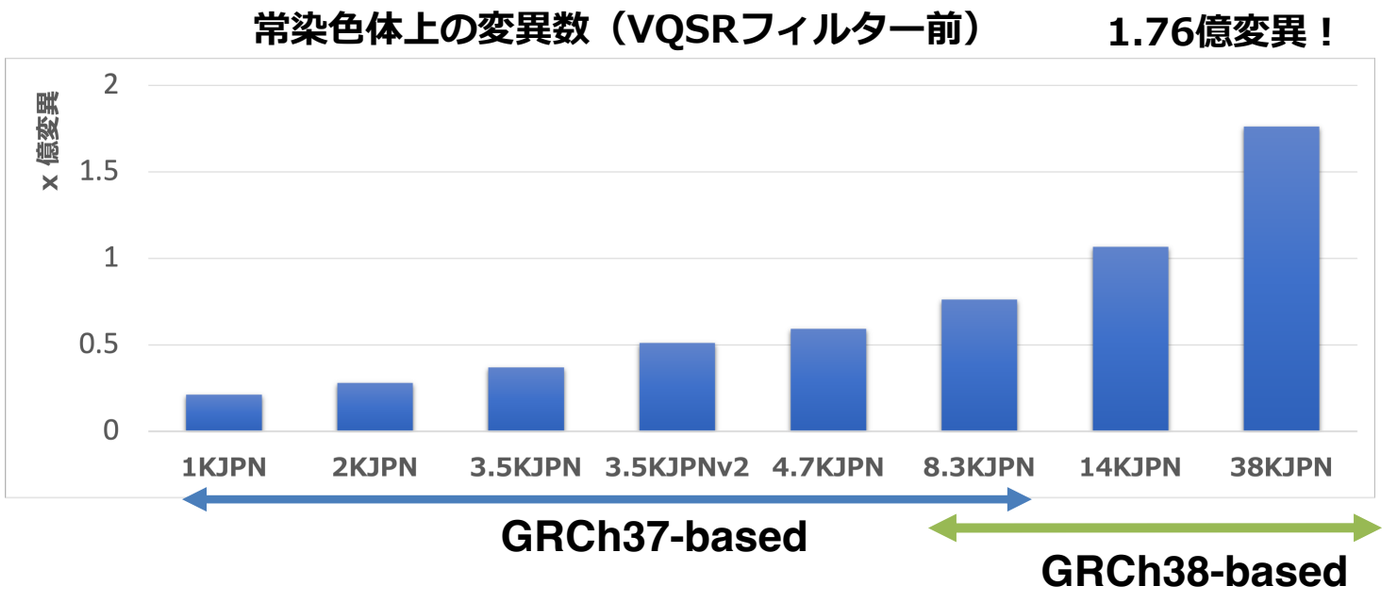
50,345検体
(50KJPN レポジトリ)

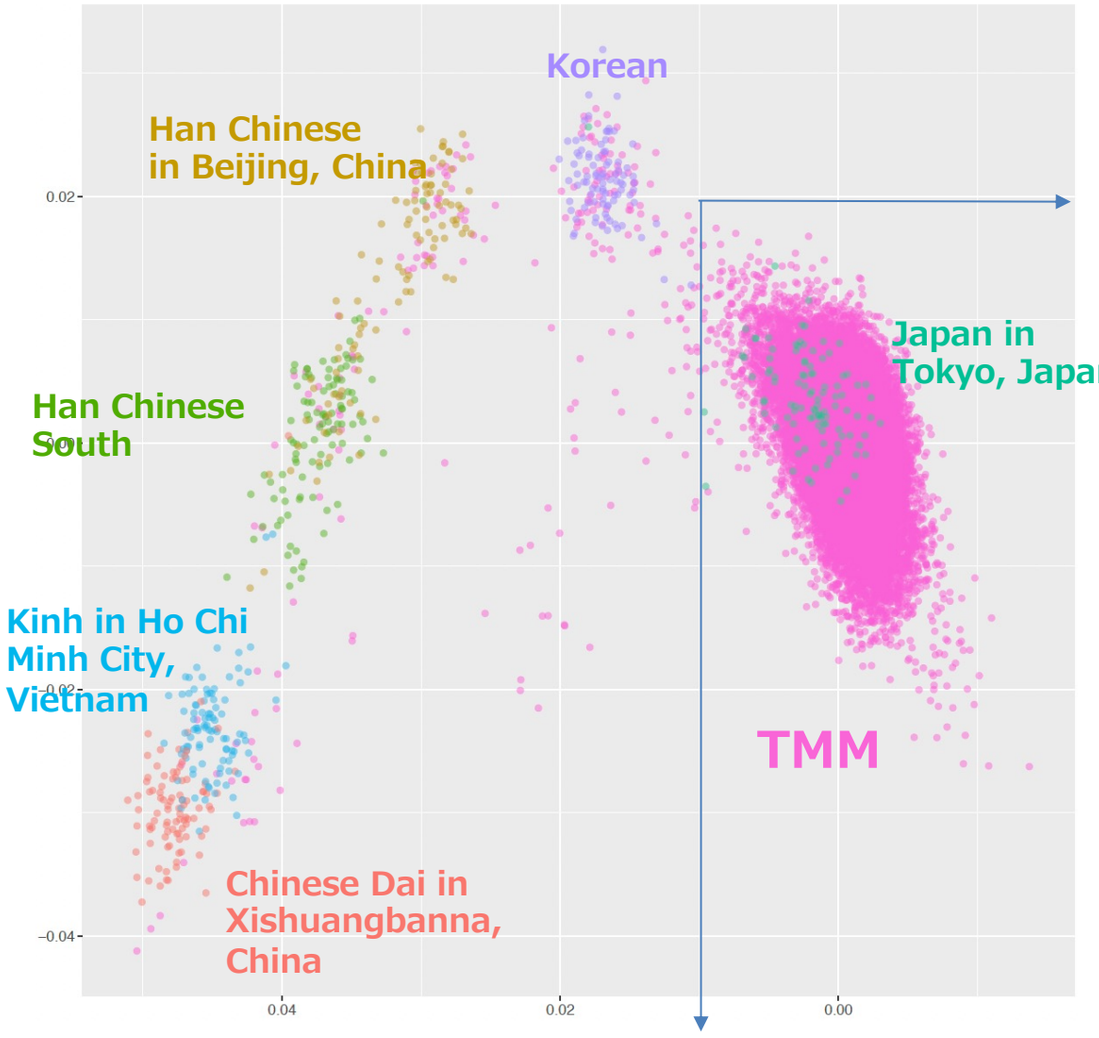
近親排除: -11,403検体

38,942検体

日本人集団以外の検体除外: -220検体
右図のPCAで判断

38,722検体
(38KJPN パネル)

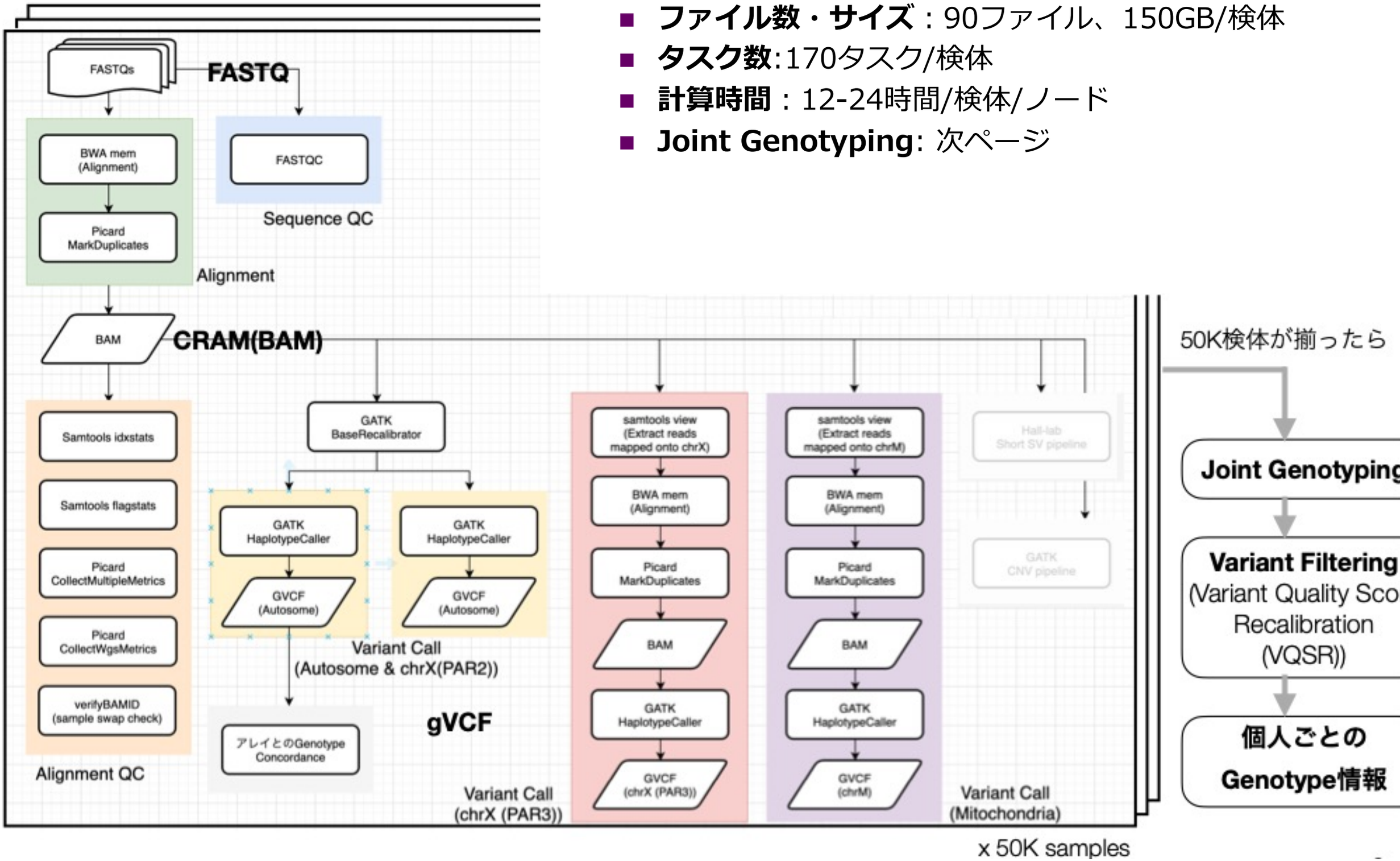




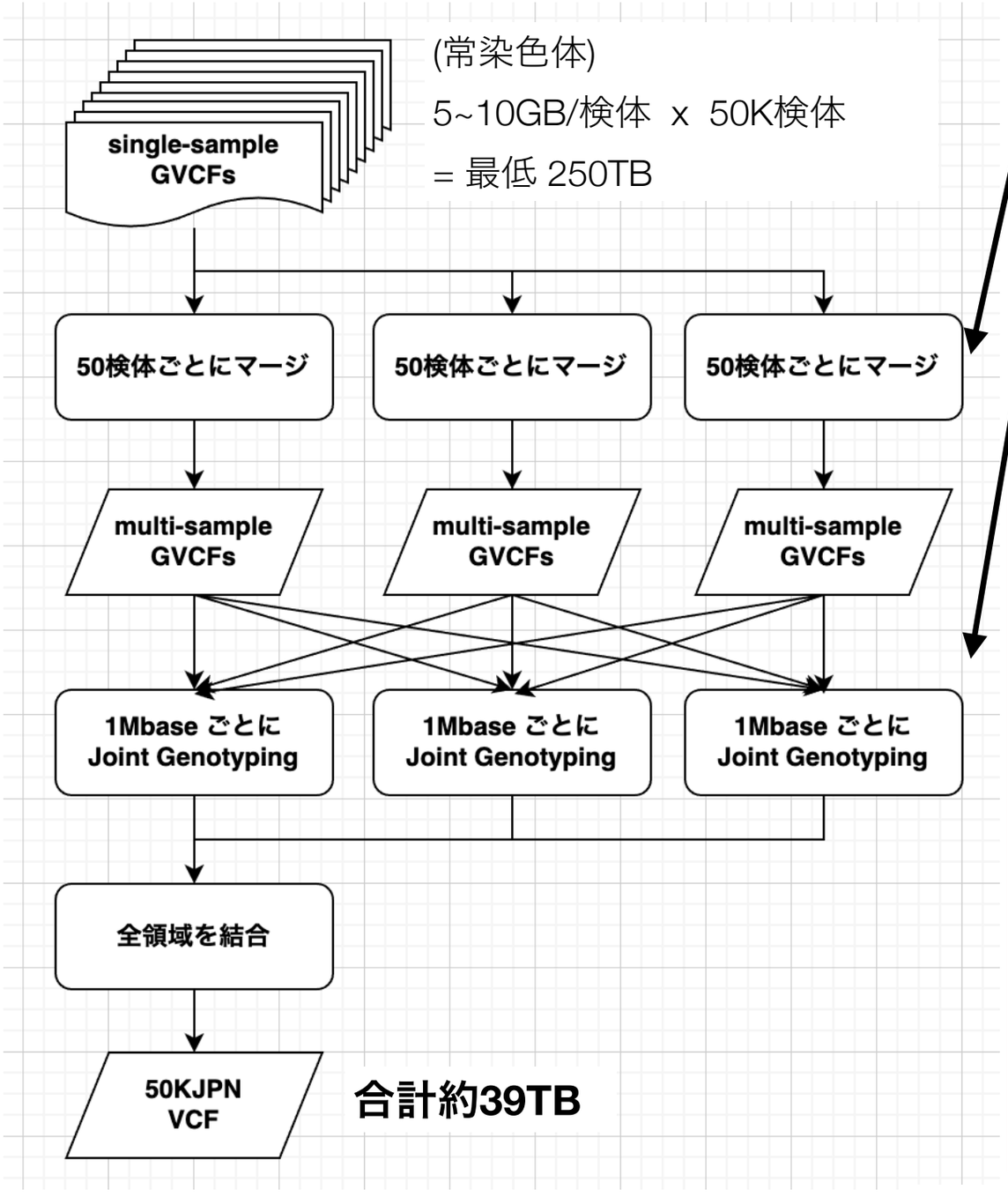
国際1000人ゲノム＋The personal genome project Korea (KPGP)で公開されているデータ、合計38,942検体と合わせてPCA

ヒト全ゲノム解析の計算量

- **ファイル数・サイズ**：90ファイル、150GB/検体
- **タスク数**:170タスク/検体
- **計算時間**：12-24時間/検体/ノード
- **Joint Genotyping**: 次ページ



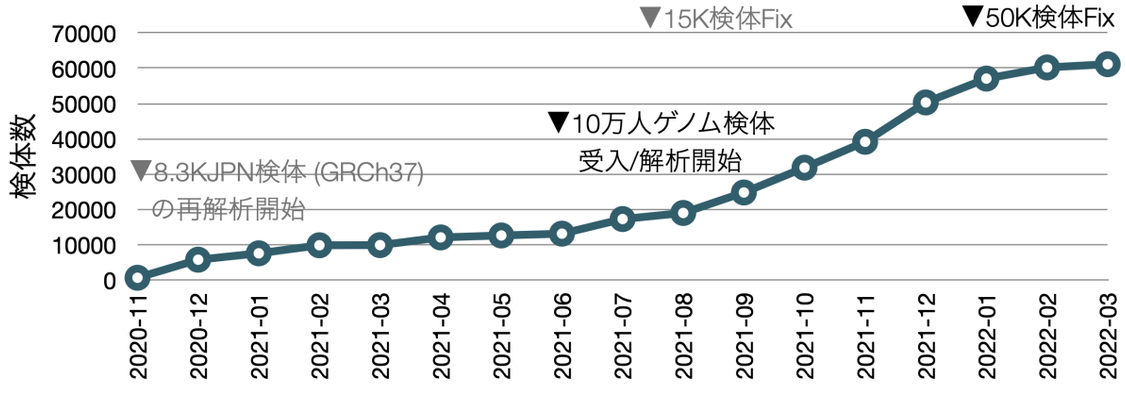
ヒト全ゲノム解析の計算量（Joint Genotyping）



50K検体分のGVCFを結合。
合計1000タスク。12時間/タスク

高速化のためゲノム領域(3Gbase) を 1Mbase に分割し、領域ごとに50K検体を処理。
合計3000タスク。24時間/タスク

- 前ページも含めて計算に必要なリソース等を合計すると
  - 最終結果を保存するストレージ
    (150GB/検体 * 50K検体) + 39TB = 7.5PB
  - タスクの数
    (170タスク/検体 * 50K検体)
    + (1000 + 3000) = 850万タスク
  - 計算時間 (並列化なしで直列に実行した場合)
    (12H * 50K) + (12H * 1000 + 24H * 3000)
    = 15.3年



実際は差分の35000検体を約6ヶ月で解析完了

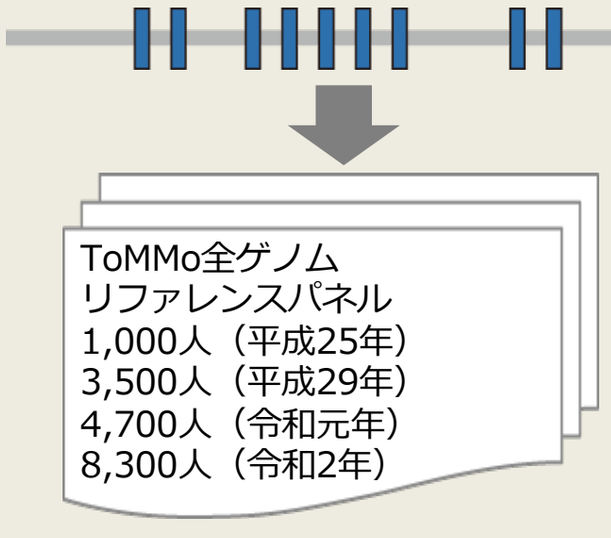
日本人向けに高度に最適化したDNAアレイの作出

ジャポニカアレイ® NEO

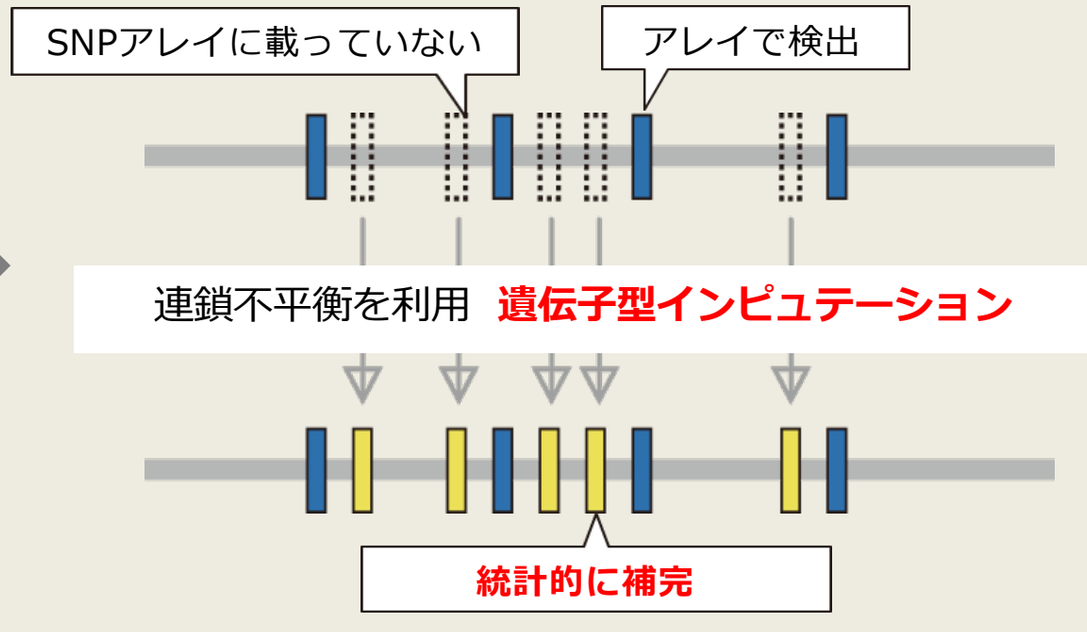
- 高品質の東北メディカル・メガバンク全ゲノム参照パネルからデザインしたもの
- SNP数を最小化しつつ疑似全ゲノム解読を可能にする
- 多くのコホート研究に活用され、個別化医療・個別化予防の普及による社会の活力向上に資することが期待される

2014年にジャポニカアレイv1、2017年にv2、
2019年に刷新版ジャポニカアレイ® NEOを発表

NGSによる全ゲノム解析（10-20万円/人）



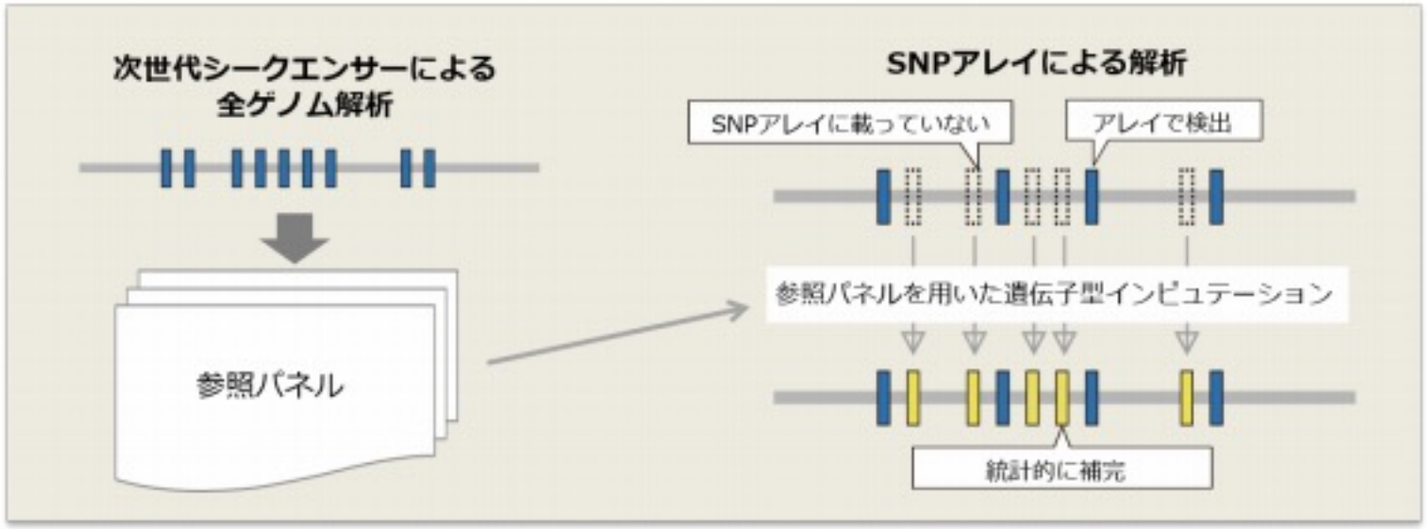
ジャポニカアレイによる解析（1万円程度/人）



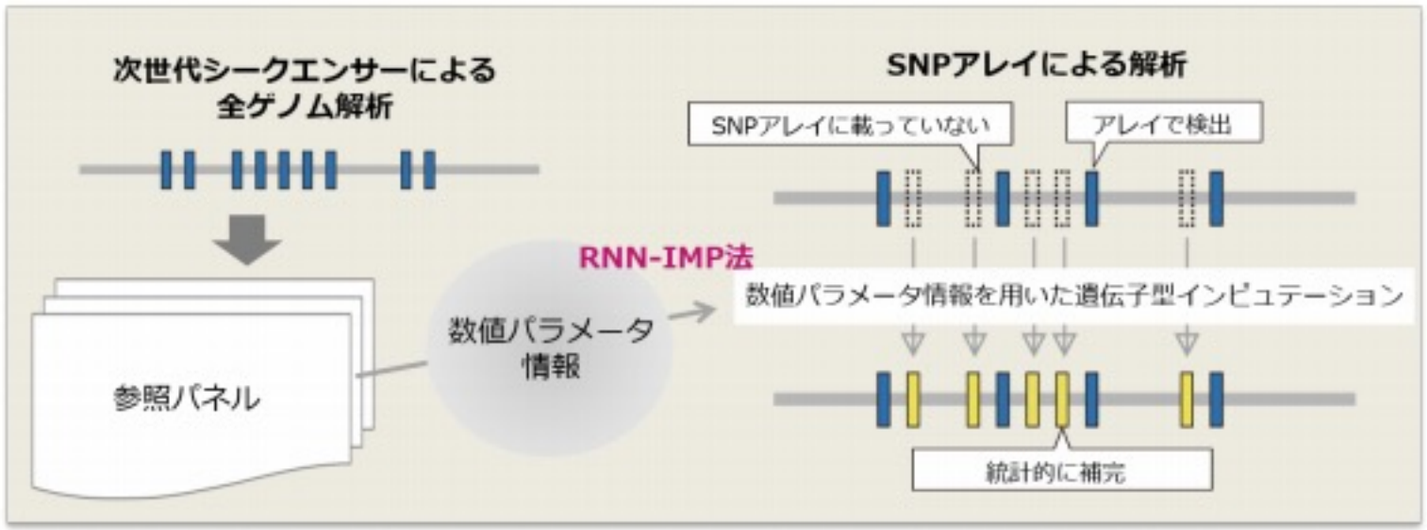
Sakurai-Yageta et al, J Biochem, 2021

AI活用での遺伝子型インピュテーション

従来の遺伝子型推定問題（個人情報が必要）



深層学習を用いたモデル化（**個人情報共有が不要**）

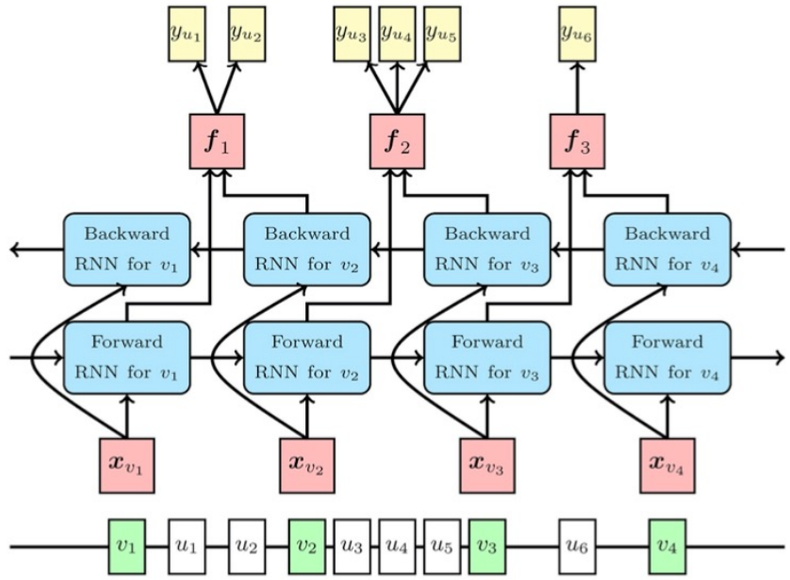


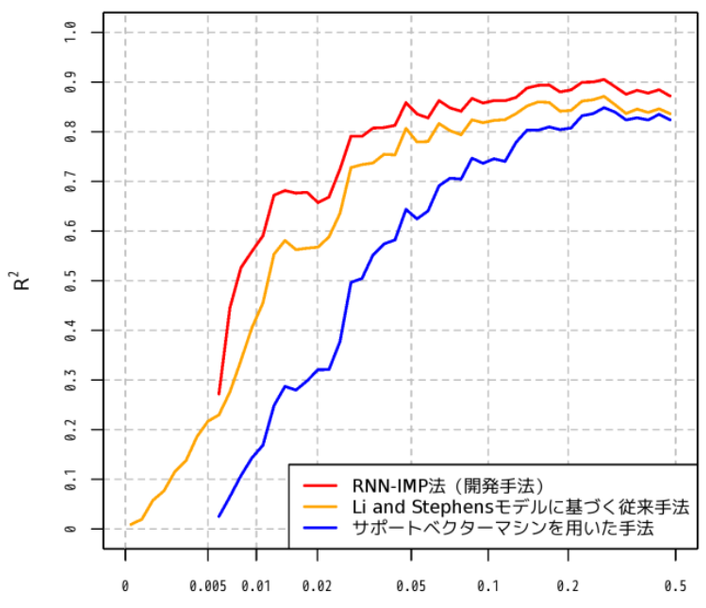
RESEARCH ARTICLE **Plos Compt Biol, 2020**

A genotype imputation method for de-identified haplotype reference information by using recurrent neural network

Kaname Kojima[1,2], Shu Tadaka[1], Fumiki Katsuoka[1], Gen Tamiya[1,2], Masayuki Yamamoto[1,3,4], Kengo Kinoshita[1,4,5,6]*

遺伝子型推定問題をRNNでモデル化することで個人情報を共有することなく推定可能に

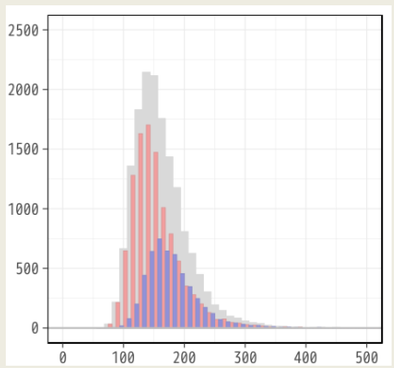




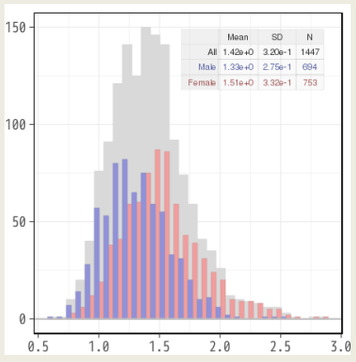
大規模メタボローム解析によるオミックス参照パネルの拡張

- 大規模メタボローム解析を実施し日本人の標準的な代謝プロファイルを高精度化
- 変わらないゲノムに対して、**動的に変化する情報として有用**
- 「日本人多層オミックス参照パネル」のメタボローム解析情報を**50,000検体に拡張・公開（jMorp 2022）**（2022年9月末拡張版公開）
- 新たな定量メタボローム解析手法の活用により、**対象化合物を拡張**
- **詳細二次調査の対象者2,900人のメタボローム解析を実施**し、各個人の代謝プロファイルの経時変化も公開
- ゲノムとメタボロームの関連解析（**MGWAS**）の規模を拡大し**多数の関連を同定**

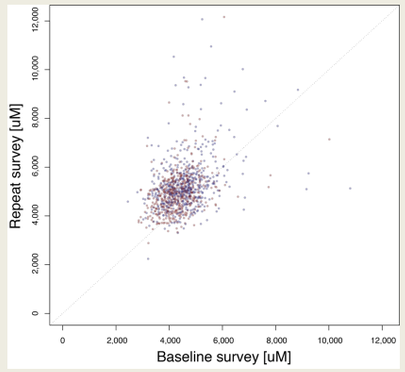
解析検体の規模を拡大
高精度代謝物分布を提供



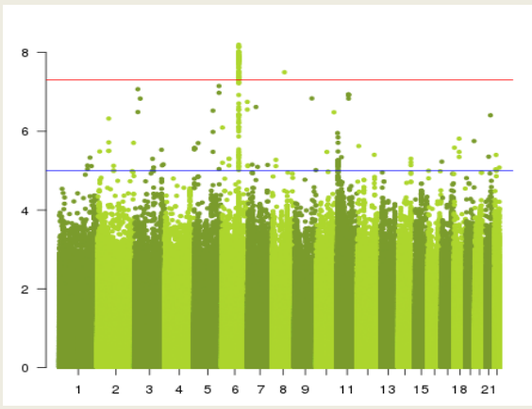
代謝物数を拡張
新規代謝物を多数公開



代謝プロファイルの
経時変化を解明



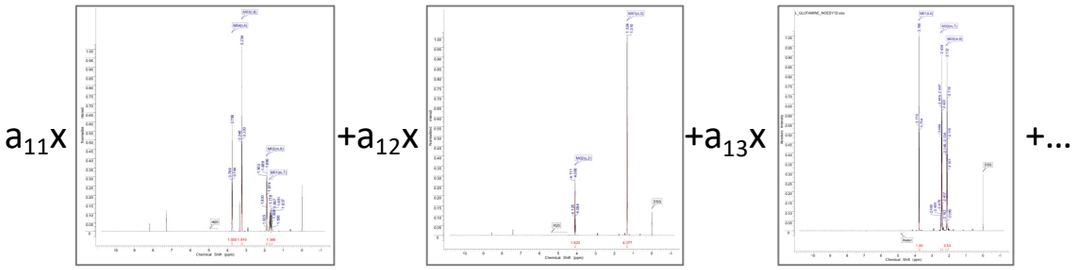
MGWAS解析の拡大による
新規関連多型の同定



国内外の研究者にオミックス参照パネル情報を提供

大規模メタボロームの実現の裏にもAI

NMR測定値は定量性が高い ↔ 1検体1化合物毎の専門家による手作業での定量化が必要

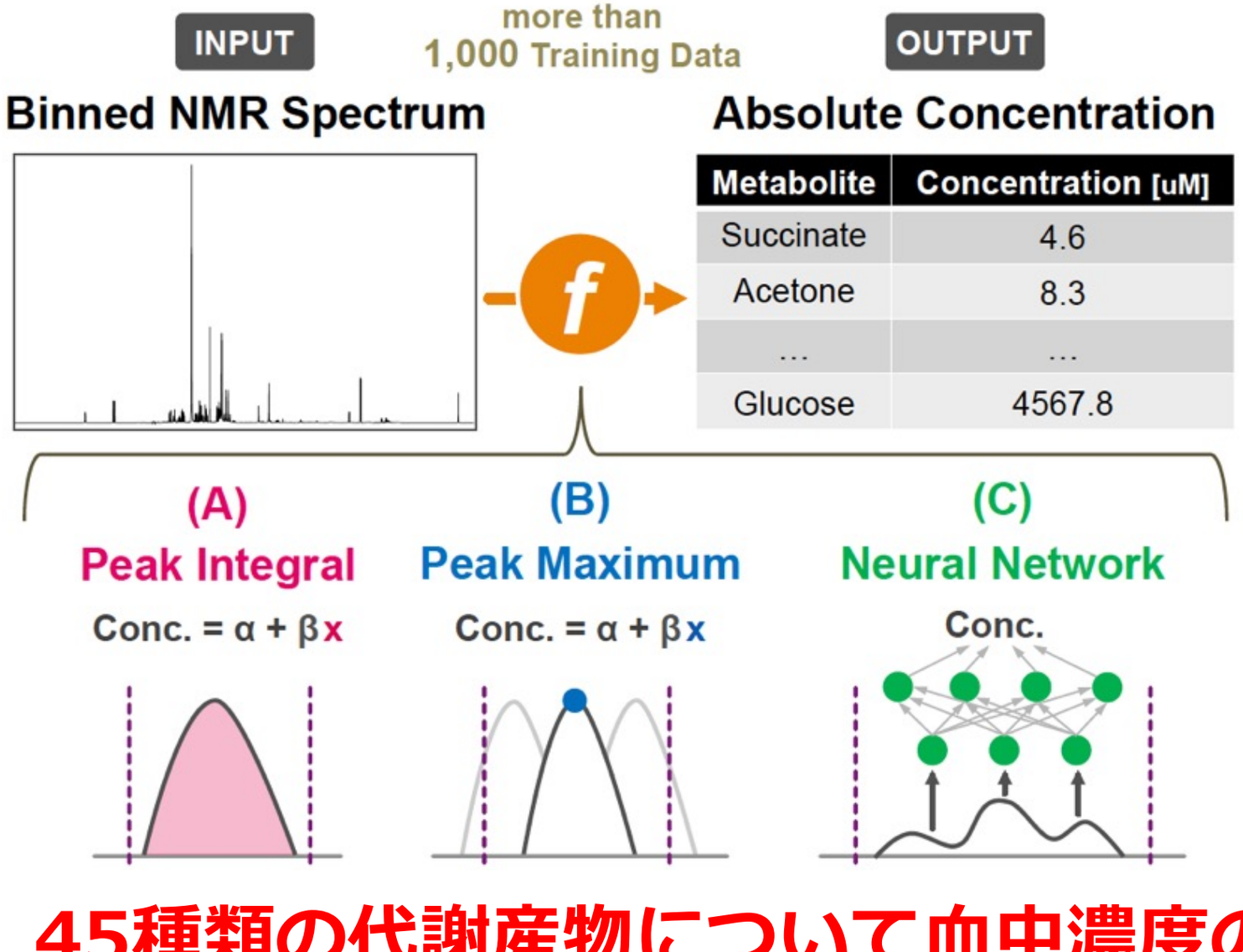


課題

これらの総和スペクトルを元に
逆問題を解く作業

500 人分40 種の代謝物の定量化：
専門家一人(専従) x 2~3 ヶ月程度

大規模化のための自動定量開発が必要



Metabolite	R2 Score	Reliability	Approach
Succinate	0.559	☆☆☆	Neural Network
Acetone	0.989	★★★	Peak Integral
3-Hydroxyisobutyrate	0.945	★★★	Peak Integral
Formate	0.779	★★☆	Peak Integral
2-Aminobutyrate	0.824	★★☆	Peak Integral
3-Methyl-2-oxovalerate	0.901	★★★	Peak Integral
Methionine	0.721	★★☆	Neural Network
2-Oxoisocaproate	0.949	★★★	Peak Integral
Creatine	0.955	★★★	Neural Network
2-Hydroxybutyrate	0.964	★★★	Peak Integral
Acetate	0.952	★★★	Peak Integral
Tryptophan	0.772	★★☆	Peak Integral
Carnitine	0.616	★☆☆	Neural Network
Cysteine	0.636	★☆☆	Peak Integral
Glutamate	0.730	★★☆	Peak Integral
Creatinine	0.935	★★★	Peak Integral
Asparagine	0.605	★☆☆	Peak Maximum
Isoleucine	0.947	★★★	Peak Integral
Phenylalanine	0.909	★★★	Peak Integral
Arginine	0.550	☆☆☆	Peak Integral
Ornithine	0.867	★★☆	Peak Maximum
Tyrosine	0.955	★★★	Peak Integral
Pyruvate	0.960	★★★	Peak Integral
Glycerol	0.832	★★☆	Neural Network
Histidine	0.837	★★☆	Peak Maximum
3-Hydroxybutyrate	0.987	★★★	Peak Integral
Leucine	0.870	★★☆	Peak Integral
Serine	0.845	★★☆	Peak Integral
Lysine	0.826	★★☆	Peak Integral
Threonine	0.928	★★★	Peak Integral
Proline	0.905	★★★	Peak Integral
Valine	0.971	★★★	Neural Network
Glycine	0.941	★★★	Peak Integral
Alanine	0.982	★★★	Neural Network
Glutamine	0.783	★★☆	Neural Network
Lactate	0.975	★★★	Peak Integral
Glucose	0.962	★★★	Peak Integral
3-methyl-2-oxobutyric acid	0.807	★★☆	Peak Integral
Betaine	0.942	★★★	Peak Maximum
Caffeine	0.979	★★★	Peak Integral
Citrate	0.925	★★★	Peak Integral
Hypoxanthine	0.995	★★★	Peak Integral
Inosine	0.981	★★★	Peak Maximum
N,N-dimethylglycine	0.718	★★☆	Peak Maximum
Uridine	0.676	★☆☆	Peak Maximum

45種類の代謝産物について血中濃度の定量自動化を実現

データ大規模化のためのデータ共有への課題

- ゲノムデータや健康調査情報は個人情報
- プライバシー保護の観点から**セキュリティ**確保が重要

相反する要件

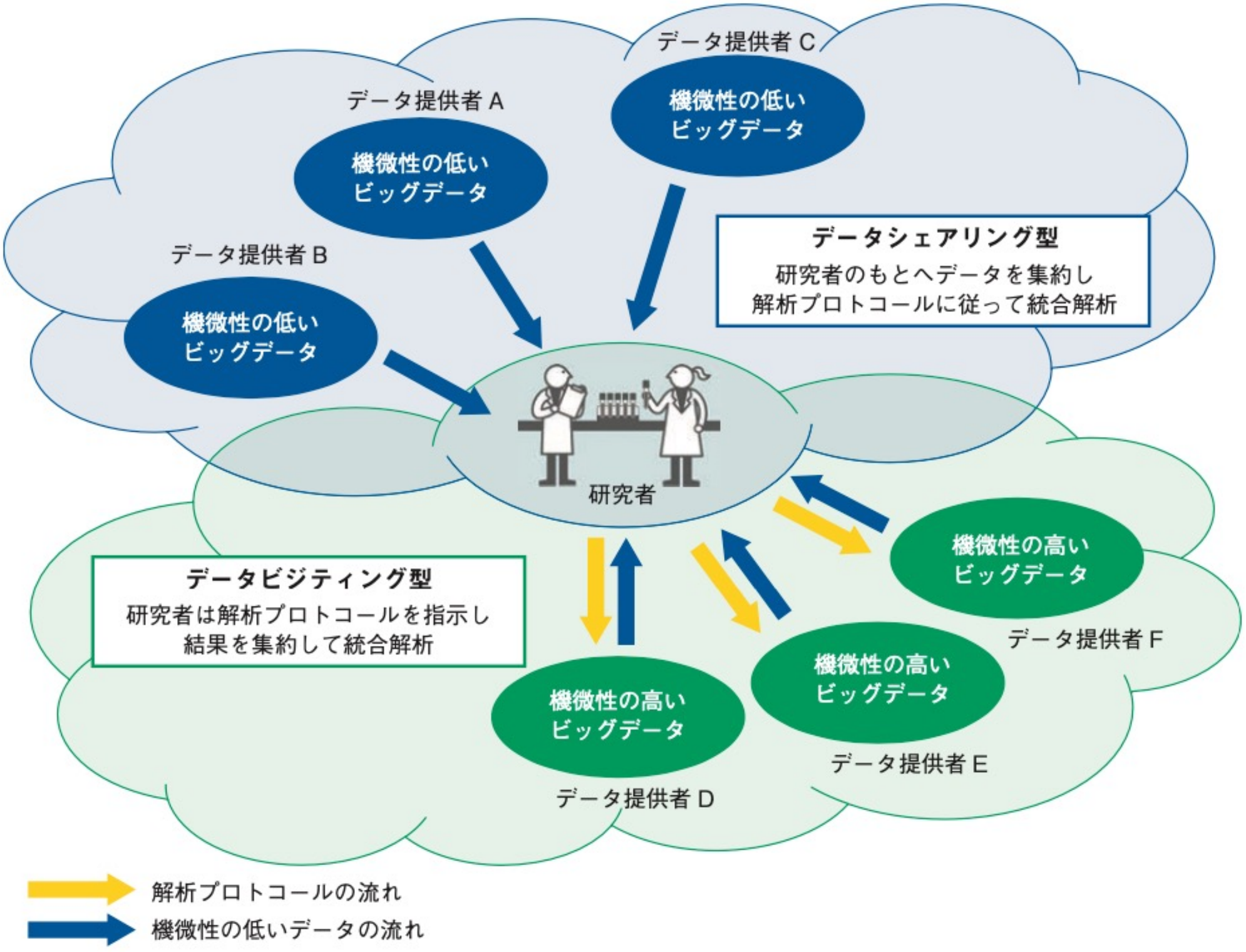
- オールジャパンでの解析に向けた**データ共有**が必須

ToMMoデータのセキュリティ分類

セキュリティ分類	情報の種類
ストロング	個人の**同定につながる可能性がある**と考えられる情報 - 例）**個人毎のゲノム変異情報、配列情報の生データ等**
スタンダード	単独では個人の**同定につながる可能性が低い**と考えられる情報 - 年齢、性別の基本情報 - 検体検査情報、調査票情報、罹患歴等の健康調査情報等 ※組み合わせにより、個人の同定の可能性が高まる属性情報の利用の場合にはストロングとする場合がある。 ※罹患歴について、コホート参加者で数名に限定される病気については、個人の同定の可能性を低くするため、属性情報をより大くくりにするか不明とする。
セミオープン	個人の**同定につながる可能性がほとんどない**と考えられる情報だが、一定の配慮が必要な場合があるデータ - パネルのジェノタイプ頻度情報
オープン	個人の**同定につながる可能性**がないと考えられる、大まかな統計情報 - 全ゲノムリファレンスパネルの頻度情報

機微性の高いデータの共有モデル

実験医学、木下賢吾（2021）

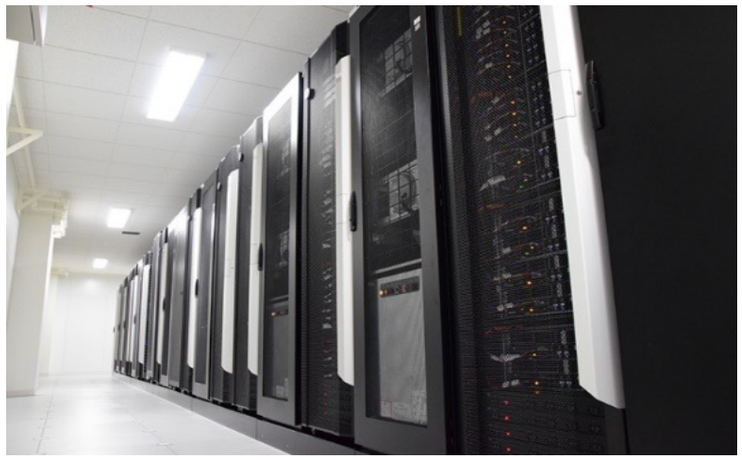


From data-sharing to data-visiting

Data-Visiting modelの実装と全国展開

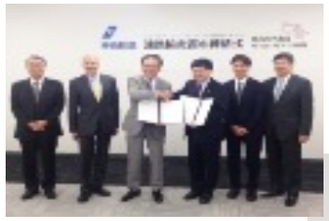
- dbTMM の膨大なデータと計算資源に高度なセキュリティを保って遠隔地からアクセス
- 幅広いデータシェアリングに貢献

東北メディカル・メガバンク機構のスーパーコンピュータ



AMED
ゲノム医療研究支援機能

- 高度セキュリティエリアからスパコンへのVPN回線によるリモートアクセスの運用
- ゲノム解析データはじめ多様なデータを安全に共有

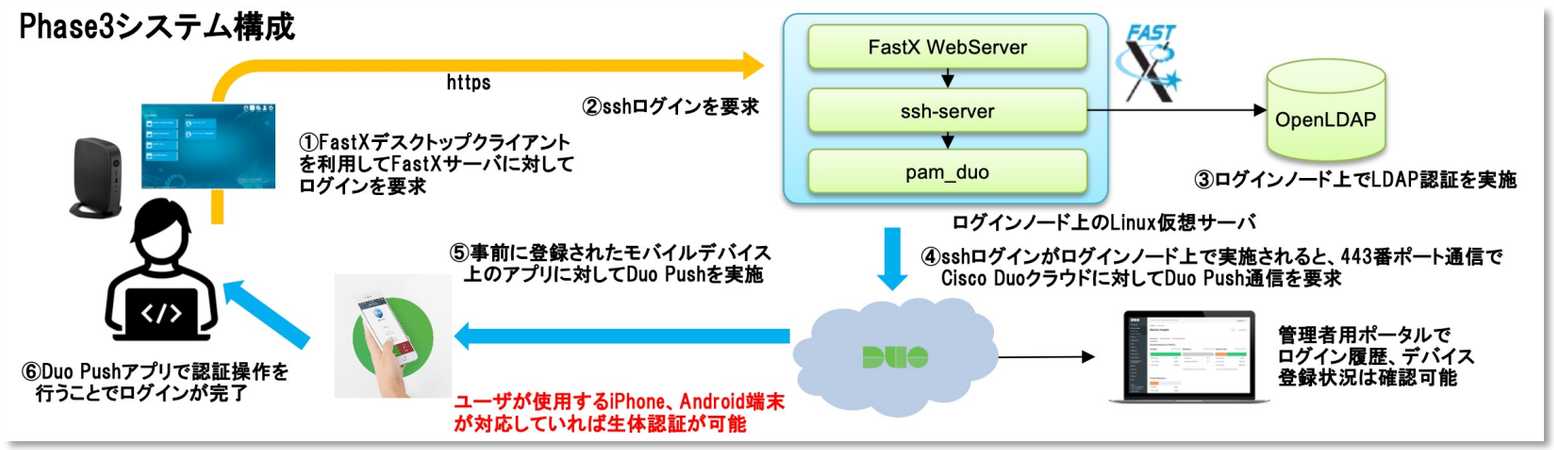


ToMMo スパコン

必要要件と特徴

- **セキュリティと利便性の両立**
  - ・FIDO認証＋専用シンクライアントによるアクセス管理
  - ・UnitA: インターネットアクセス可だが機微性の高いデータ不可
  - ・UnitB: 遠隔セキュリティエリアからアクセス
    限定的な機微性の高いデータ可
  - ・UnitC: 原則としてToMMo内限定、全てのデータ解析可
- **大規模ゲノム解析に耐える性能**（CPUよりIO)
  - ・CPUはコスト的に美味しいところをできるだけ多く
  - ・Diskは並列ファイルシステムLustreの国内最大サイト
- **ゲノムだけで無く多様な解析に使える一般的なアーキテクチャ**

	Phase1	Phase2	Phase3
Launch Time	2014.4	2018.4	2022.4
CPU (core)	16,480	7,200（+V100x24）	14,080(+A100x24)
Disk (PB)	18	27	50(+10 Scality)



本日の話題

- ゲノムデータの可能性と限界
- 個別化医療への課題と展望
- 東北メディカル・メガバンク計画によるデータ創生と活用
- 機微性の高いデータ共有に向けて
- ToMMoスーパーコンピュターシステム
- **まとめ**

未来型医療創成センター（INGEM）を通じた社会実装

・指定国立大学の指定の際に設立した４つの「東北大世界トップレベル研究拠点」の１つ
・10部局※が参画し、未来型医療拠点（個別化予防・医療）を構築

※東北大学病院、医学系研究科、加齢医学研究所、情報科学研究科、歯学研究科　薬学研究科、医工学研究科、東北メディカル・メガバンク機構、工学研究科、生命科学研究科

・ToMMoの住民型バイオバンク＋病院の疾患型バイオバンクの連携
・情報科学研究科も加わり情報インフラ、解析も協力に実施
・次世代シークエンサ、クライオ電顕など最先端解析機器を所有
・多分野の力を結集した研究拠点形成で世界最先端研究を実現

同一地域の比較対照
設備・技術の効率的利用

東北大学病院

疾患型バイオバンク

- 宮城県第三次医療拠点
- がんゲノム医療中核拠点
- 臨床開発拠点CRIETO

バイオバンク管理技術・
データベース・ゲノム解析・
NMR・質量分析など

コホート参加者の疾患発症
前向き研究のアウトカム
発症前後の試料・情報

東北メディカル・メガバンク機構

住民型バイオバンク

15万人の試料・情報
- 地域住民コホート研究
- 三世代コホート研究

INGEMは疾患型バイオバンクと住民型バイオバンクとを統合することで
新たな個別化医療を創成する世界トップレベルの研究拠点を目指す

まとめ

- **ゲノム情報の可能性と限界**
  ゲノム情報の持つ可能性は大きい
  ゲノムデータが身近になってきている
  一方で表現型との関係のエビデンスが不足し実用化に課題も残る
- **東北メディカル・メガバンク計画によるデータ創生**
  良質なデータをコホート調査で収集
  日本人集団でのエビデンスの蓄積
  日本のゲノム医療の**インフラ**としてデータ提供を推進
  ジャポニカアレイをキーテクノロジーとした未来型医療への挑戦
  東北大学病院と連携した未来型医療創成センター(INGEM)の創設

Thanks to

他に23名の室長や７名の地域支援センター長合計約380名程度のスタッフ（GMRC / TCFを含む）

GMRC: genome medical research coordinator　TCF: ToMMo clinical fellow

Special Thanks to

Japonica Array NEO: 櫻井美佳

38KJPN: 田高周、川嶋順子、岡村容伸、勝岡史城

RNNインピュテーション: 小島要

スパコン: 岡村容伸、田高周